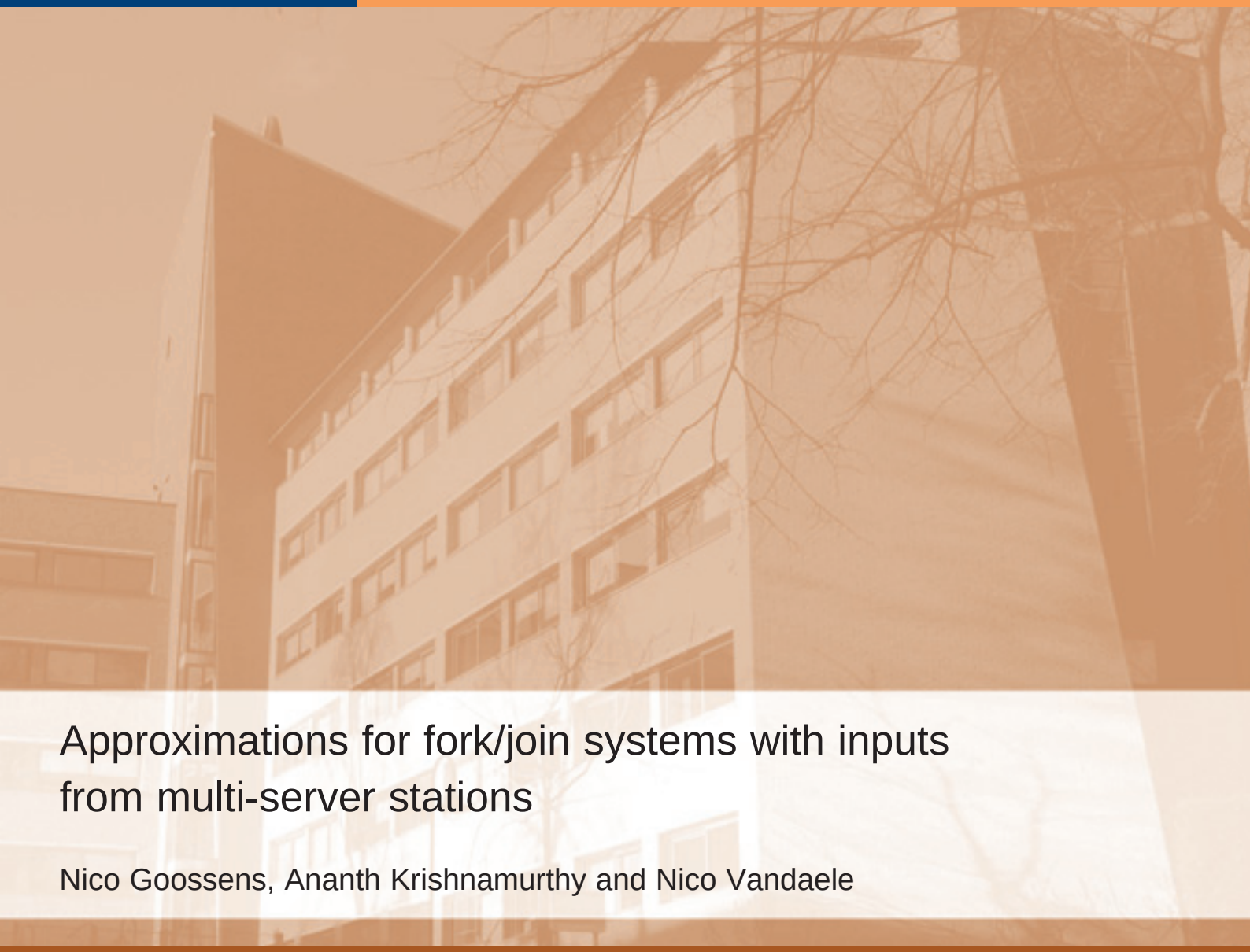
KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics



Approximations for fork/join systems with inputs from multi-server stations

Nico Goossens, Ananth Krishnamurthy and Nico Vandaele

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0737

Approximations for Fork/Join Systems with Inputs from Multi-Server Stations

Nico Goossens
OM Partners n.v.,
Koralenhoeve 23 2160 Wommelgem Antwerpen, Belgium.
Email: ngoossens@ompartners.com
*Fax:*32-3-650.22.90

Ananth Krishnamurthy[1]
Department of Decision Sciences and Engineering Systems,
Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY-12180, USA.
Email: krisha@rpi.edu
*Phone:*1-518-276-2958, *Fax:* 1-518-276-8227

Nico Vandaele[2]
Faculty of Applied Economics
Katholieke Universiteit Leuven, Campus Kortrijk, Belgium.
Email: Nico.Vandaele@kuleuven-kortrijk.be

Abstract

Fork/join stations are commonly used to model synchronization constraints in queuing network models of computer and manufacturing systems. This paper presents an exact analysis of a fork/join station in a closed queuing network with inputs from multi-server stations with two-phase Coxian service distributions. The underlying queue length process is analyzed exactly to determine performance measures such as throughput, and distributions of the queue length at the fork/join station. By choosing suitable parameters for the two-phase Coxian distributions, the effect of variability in inputs on system performance is studied. The study reveals that for several system configurations, analysis of the simpler system with exponential inputs provides efficient approximations for performance measures. Both, the exact analysis and the simple approximations of fork/join systems constitute useful building blocks for developing efficient methods for analyzing large queuing networks with fork/join stations.

Keywords: queueing, fork/join, synchronization, assembly systems, closed queuing networks

[1]Corresponding author: Ananth Krishnamurthy (E-mail: krisha@rpi.edu, Phone: 1-518-276-2958)
[2]Formerly at the University of Antwerp, Department MTT.

1 Introduction

As manufacturing and computer systems become more complex, executing operations in parallel is seen as a way to improve efficiencies and responsiveness. Therefore, it is important to understand the effect of synchronization constraints imposed on parallel operations on overall system performance. Queuing network models with fork/join constraints have been used in a variety of applications to evaluate the effect of synchronization constraints on system performance. In queuing models of fabrication/assembly systems, fork/join stations model synchronization constraints prior to assembly operations (Harrison [12], Baynat and Dallery [4], Rao and Suri [26], and de Boeck [8]). In computer systems analysis, queuing networks with fork/join stations have been studied in the context of parallel processing, database concurrency control, and communication protocols (Baccelli et al. [3], Varki [30], Prabhakar et al. [23]).

The fork/join station used to model synchronization constraints in most applications typically consists of two or more input buffers. A fork operation generates arrivals of entities to the input buffers of the fork/join station. The entities arrive at each input buffer according to a random process and if the required entities are available in each input buffer, an entity is removed from each buffer and joined together. The joined entity is released from the fork/join station instantaneously. The performance measures of interest include synchronization delays, queue length distributions at the different input buffers, and station throughput. Earlier works on the analysis of fork/join stations investigate stability conditions and derive performance estimates when the inputs to the individual buffers are Poisson processes (Bhat [5], Harrison [12], Som et al. [27]). Subsequent studies have extended the analysis to systems where the inter-arrival time distributions of the inputs to each buffer have phase type distributions (Takahashi et al. [29]). All these studies assume that the arrival process to the fork/join station is independent of the buffer contents at the station. However, when the fork/join station is part of a closed queuing network, the rate of arrivals to the fork/join station may be self regulating or a function of the contents of its input buffers. The effect of such arrival processes on the performance a fork/join station has received interest in recent years (Krishnamurthy et al. [18], [17], Krishnamurthy and Suri [16], Goossens et al. [11], Baynat and Dallery [4]). All these studies assume that the input to each buffer is from a closed network consisting of a station with fixed or variable service rates. In particular, Goossens et al. [11] models a fork/join station in with inputs from a closed network with multi-server stations having exponentially distributed service times. The study reveals that the performance of the fork/join station could be significantly different from those with

inputs from single servers.

This research extends the findings in Goossens et al. [11] and studies the effect of variability on fork/join stations with inputs from multi-server stations. In particular, an exact analysis of fork/join stations with inputs from finite population sub-networks with multi-server stations is conducted assuming that the service times at the stations have a two-phase Coxian distribution. The choice of two-phase Coxian distribution permits analysis of a wide class of variability in input processes. The analysis reveals that the effect of variability in inputs is significant in only certain regions of the input parameter space. This property could be very useful when designing systems that need to be robust to variability in inputs. Additionally, the potential of using simple approximations based on exponential inputs are explored. Numerical studies indicate that these approximations could save computational effort and yet predict performance measures that are within 5% of their true values.

The remainder of this paper is organized as follows. Section 2 provides a summary of the literature to date on the analysis of fork/join stations. Section 3 defines the model of the fork/join station and summarizes the analysis approach. Section 4 describes the analysis of the queue length process, and Section 5 investigates the effect of variability in inputs on key performance measures. Section 6 investigates the use of systems with exponential inputs to provide quick and efficient approximations for more general systems. Section 7 provides insights with respect to the variability in the inter-departure times from the fork/join station and Section 8 presents the conclusions.

2 Literature Review

Fork/join stations have been extensively studied in the context of queuing models of computer and manufacturing systems. Harrison [12] and Latouche [20] analyze stability conditions for fork/join stations and conclude that enforcing specific bounds on the size of the input buffers is one way to guarantee stability. Bhat [5] analyzes a fork/join station with finite buffers assuming Poisson inputs and derives expressions for the queue length distributions at the input buffers. Baccelli and Makowski [2], Nelson and Tantawi [22], Kumar and Shorey [19], Bonomi [6], Liu and Perros [21] analyze fork/join stations with Poisson inputs and evaluate the performance of fork/join stations under different settings. Knessl [14] and Varma and Markowski [31] approximate the queue length distribution under heavy traffic limits using diffusion approximations. Prabhakar et al. [23] study the departure process of

fork/join stations with Poisson inputs under limiting conditions. Som et al. [27] and Takahashi et al. [28] study the departure process from a fork/join station with Poisson inputs. They derive expressions for the marginal distribution of the inter-departure times from the fork/join station. Takahashi et al. [29] subsequently extends the analysis for systems where the inputs have phase type distributions. Ko and Serfozo [15] develop bounds and approximate expressions for evaluating the mean response time and queue length distribution at fork/join stations with inputs from multi-server stations with exponential service times.

More recently, studies on fork/join stations have focused on a variant of the systems studied above, namely, systems where the inputs are from finite populations. For instance, when the fork/join station is part of a larger closed queuing network, then once the content in the input buffer reaches a certain level, the arrival process shuts down temporarily. Varki [30] uses mean value analysis to study fork/join station performance in a closed queuing network under exponential settings. Krishnamurthy et al. [18] evaluate performance measures of fork/join stations with inputs from finite population subnetwork with stations having 2-phase Coxian distributions. Subsequent analysis by Ramakrishnan and Krishnamurthy [25] proposes approximations for fork/join stations with two or more inputs. However, all these prior works assume that inputs to the synchronization station are from single server stations. Goossens et al. [11] models a fork/join station with inputs from multi-server stations with exponentially distributed service times. This paper compliments prior efforts by Baynat and Dallery [4], and Di Mascolo et al. [9], Goossens et al. [11] by studying the effect of variability in fork/join stations where the inputs to the buffers are from finite population sub-networks with multi-server stations. The details are presented in the subsequent sections.

3 System Description

Figure 1 represents a fork/join station, J with inputs from two multi-server stations. Correspondingly, the station has two input buffers B_1 and B_2. If an entity arriving in buffer B_1 (B_2) finds input buffer B_2 (B_1) empty, it waits for the corresponding entity to arrive in input buffer B_2 (B_1). As soon as there is at least one entity in each buffer, one entity is removed from each buffer. The removed entities join together, and immediately depart from the fork/join station. As a result the content of each input buffer is reduced by one. Subsequent to departure from the fork/join station, the joined entity forks back into two entities that are routed back to station i, $i = 1, 2$ respectively, where they wait in queue (if necessary) for service. Station i consists of c_i identical parallel servers and the service time at each server

is assumed to have a two-phase Coxian distribution. Upon completion of service at station i, the entity waits in buffer B_i on a first come first served basis. There is a finite population of size K_i for the entity of type i, and it is assumed that $K_i \geq c_i, i = 1, 2$. Consequently, the number of entities in input buffer B_i and at the corresponding servers at station i always sum up to K_i , $i = 1, 2$, and the arrival process to buffer B_i shuts down temporarily when there are K_i units in buffer B_i, and resumes following the next departure from the fork/join station.

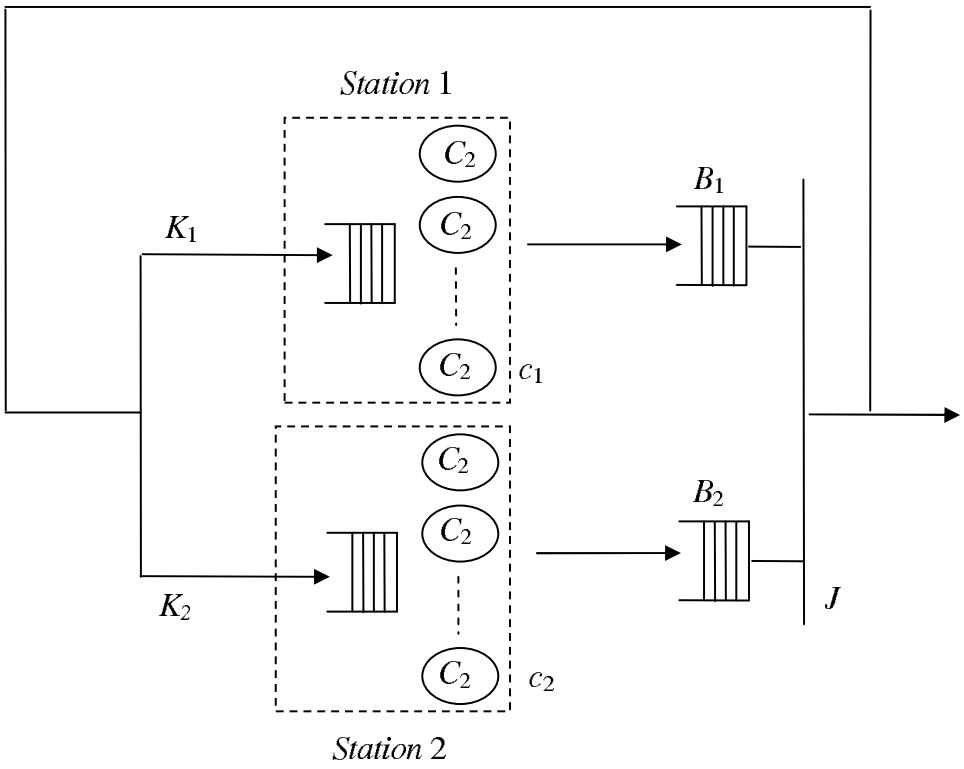


Figure 1: Fork/join station with inputs from multiple servers

A two-phase Coxian distribution is chosen for the service time at each server in order to be able to analyze the effect of the mean and variability in service times on the key performance measures of the fork/join station. At a server in station $i, i = 1, 2$, the service process for an entity first includes an exponential phase characterized by the parameter μ_{i1}. Subsequently, with a probability θ_i, the service could involve another exponential phase characterized by the parameter μ_{i2}. Alternatively, with a probability $1 - \theta_i$, the service is complete upon completion of the first exponential phase of service. If $G_i(t), i = 1, 2$ denote the distribution functions of the service times at the two stations, then

$$G_i(t) = 1 - C_{i1}e^{-\mu_{i1}t} - C_{i2}e^{-\mu_{i2}t} \text{ for } t \geq 0 \tag{1}$$

where

$$C_{i1} = \frac{\mu_{i1}\left(1-\theta_i\right)-\mu_{i2}}{\mu_{i1}-\mu_{i2}} \qquad \text{and} \qquad C_{i2} = 1 - C_{i1}, \qquad \text{with} \qquad \mu_{i1} \neq \mu_{i2}. \tag{2}$$

This implies that the mean, μ_i^{-1}, and SCV, $c_{S,i}^2$, of the service time at station i, $i = 1, 2$ are given by:

$$\frac{1}{\mu_i} = \frac{1}{\mu_{i1}} + \frac{\theta_i}{\mu_{i2}} \tag{3}$$

$$c_{S,1}^2 = 1 - \frac{2\theta_i\mu_{i1}\left(\mu_{i2}-\mu_{i1}\left(1-\theta_i\right)\right)}{\left(\theta_i\mu_{i1}+\mu_{i2}\right)^2} \tag{4}$$

Note that, the parameters, μ_{i1}, μ_{i2} and $\theta_i, i = 1, 2$ can be set using principles suggested in Altiok [1] to model service times that that have a finite positive means and SCVs in the range $[0.5, \infty)$. If information about only mean and SCV of service times are known, then one could use this information to fit a two-phase Coxian distribution for the purpose of analysis. In this case, an additional condition of balanced means (i.e. $\frac{1}{\mu_{i1}} = \frac{\theta_{i1}}{\mu_{i2}}, i = 1, 2$) is usually assumed to completely characterize the service time distributions at each station. Table 1 summarizes the main notation.

Symbol	Description
K_i	Size of the finite population from which arrivals occur to station $i, i = 1, 2$
c_i	Number of servers at station $i, i = 1, 2$
μ_i^{-1}	Mean service time at a server at station $i, i = 1, 2$
$c_{S,i}^2$	SCV of the service time at a server at station $i, i = 1, 2$
$(\mu_{i1}, \mu_{i2}, \theta_i)$	Parameters of the two-phase Coxian distribution for service times at station $i, i = 1, 2$
λ_D	Throughput from the fork/join station
$E(L_i)$	Mean queue length at buffer $B_i, i = 1, 2$
c_D^2	SCV of the inter-departure times from the fork/join station $i, i = 1, 2$

Table 1: Notation used in the analysis

3.1 Example Applications

Fork/join stations with such characteristics are found in queuing network models of closed multi-level fabrication/assembly systems, multi-stage kanban systems, and tandem lines with multi-server stations and finite buffers.

- The fork/join station described above can represent a synchronization station before

an assembly operation in a fabrication/assembly system [26]. In this case K_i could correspond to the fixed number of automated guided vehicles (AGVs) transporting components of type i from the fabrication sub-network (represented by the multi-server stations) to the assembly station. Entities in buffers B_1 and B_2 correspond to the fabricated parts waiting for other components required for assembly. The join operation corresponds to the kitting operation, while the fork operation corresponds to the release of free AGVs to carry the parts required for assembly. The arrival of reloaded AGVs from each fabrication sub-network could be modeled using a multi-server station with general service times that is approximated by a suitable two-phase Coxian distribution.

- As a second example, the fork/join station model could represent the synchronization constraint in a kanban control system [9]. Here the fork/join station could model the synchronization constraint between an upstream stage (represented by station 1 and the downstream stage represented by station 2, in a multi-stage kanban system. Each entity in buffer B_1 would correspond to a part with an upstream kanban attached to it, while each entity in buffer B_2 would correspond to a free kanban returning from the downstream stage, and K_1 and K_2 would be the number of kanbans in the respective stages. During the join operation a part and upstream kanban are joined with a downstream kanban and during the fork operation, the upstream kanban is sent back, while the part and downstream kanban are sent to the next manufacturing stage. The manufacturing process in each fabrication sub-network could be modeled using a multi-server station with general service time that is approximated by a suitable two-phase Coxian distribution.

- As a special case, the fork/join station model could also model blocking phenomenon between two consecutive stations in multi-server tandem lines with no buffers (Goossens [10]). The upstream station, station 1 is assumed to have c_1 servers with general service times and a buffer capacity of $K_1 = c_1$, while the downstream station, station 2 is assumed to have c_2 servers with general service times and a buffer capacity of $K_2 = c_2$. Each entity in buffer B_1 would correspond to a server that is blocked after service, while each entity in buffer B_2 would correspond to an starved server in station 2. Whenever, there are c_1 entities in buffer B_1, all servers at station 1 are blocked and when there are c_2 entities in buffer B_2, all servers at station 2 are starved. Clearly, one cannot have blocked servers in station 1 (i.e. entities in buffer B_1 when there are starved servers in station 2 (i.e entities in buffer B_2). If the general service times are approximated by suitable two-phase Coxian distributions, the fork/join station precisely models the

dynamics between two consecutive stations in the tandem line and could be used as a building block for analysis of longer lines as illustrated in Goossens [10].

3.2 Overall Approach

For the fork/join station described above, the goal is to compute the throughput λ_D, and the mean queue lengths $E(L_i), i = 1, 2$ at the buffers $B_i, i = 1, 2$. These are determined by conducting an exact analysis of the underlying queue length process of the fork/join station. By defining a suitable state space, the queue length process is analyzed as a continuous time Markov chain. Using the solution to the Markov chain, numerical studies are conducted to study the effect of variability on key performance measures. Subsequently, approximations are proposed based on simpler systems with exponential inputs. A detailed experiment is conducted to quantify the accuracy of such approximations throughout the design space. Finally, some insights with respect to the variability in inter-departure times, c_D^2 from the fork/join station are provided.

4 Analysis of Queue Length Processes

This section describes the exact analysis of the queue length process of the fork/join station. The queue length process is analyzed as a continuous time Markov process, and the underlying Markov chain is solved to obtain the steady state probability distributions. From these probability distributions, performance measures such as the throughput λ_D and mean queue lengths $E(L_1)$ and $E(L_2)$ at buffers B_1 and B_2 respectively are estimated. The details are given below.

Referring to Figure 1, let $N_1(t)$ and $N_2(t)$ denote the number of units in buffers B_1 and B_2 respectively at time t. If at some time t, $N_i(t) = k_i, i = 1, 2$, then the remaining $K_i - k_i$ units are at station i and $min(K_i - k_i, c_i)$ servers at the station are busy at time t. If $N_i(t) = K_i$, then all servers at station i are idle and the arrival process to buffer B_i temporarily shuts down. The operational characteristics of the fork/join station imply that buffers B_1 and B_2 cannot be both non-empty simultaneously, i.e., if $N_1(t) > 0$, then $N_2(t) = 0$, and vice versa. To completely describe the state of the system at any time t, both-the number of units in each input buffer, $N_i(t), i = 1, 2$ as well as the phases of the pending arrivals need to be considered. Recall that, if at time t, $N_i(t) = k_i, i = 1, 2$, then $min(K_i - k_i, c_i)$ servers are busy at station i. At each busy server, the service process can either be in phase 1 or phase

2. If e_{i1} and e_{i2} denote the number of servers at station $i, i = 1, 2$ with service process in phase 1 and phase 2 respectively, then $e_{i1} + e_{i2} = min(K_i - k_i, c_i)$. With these definitions the state, s, of the system is completely characterized by the tuple $(k_1, e_{11}, e_{12}, k_2, e_{21}, e_{22})$. Clearly with this enhanced description of the system state, the stochastic behavior of the system can be evaluated as a continuous time Markov chain. The state space is given by $\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3$, where

$$\begin{aligned}
\mathbb{S}_1 &= \{(k_1, e_{11}, e_{12}, k_2, e_{21}, e_{22}) : 0 < k_1 \leq K_1; k_2 = 0; e_{11} + e_{12} = min(K_1 - k_1, c_1); e_{21} + e_{22} = c_2\} \\
\mathbb{S}_2 &= \{(k_1, e_{11}, e_{12}, k_2, e_{21}, e_{22}) : 0 < k_2 \leq K_2; k_1 = 0; e_{21} + e_{22} = min(K_2 - k_2, c_2); e_{11} + e_{12} = c_1\} \\
\mathbb{S}_3 &= \{(k_1, e_{11}, e_{12}, k_2, e_{21}, e_{22}) : k_1 = 0 = k_2; e_{21} + e_{22} = c_2; e_{11} + e_{12} = c_1\}
\end{aligned}$$

It can be shown that this finite Markov chain is positive recurrent and therefore, the steady state probabilities $P(k_1, e_{11}, e_{12}, k_2, e_{21}, e_{22})$ must satisfy the following set of balance equations:

For $K_1 - c_1 < k_1 \leq K_1, k_2 = 0$ with $e_{11} + e_{12} = min(K_1 - k_1, c_1); e_{21} + e_{22} = c_2$:

$$\begin{aligned}
& (e_{11}\mu_{11} + e_{12}\mu_{12} + e_{21}\mu_{21} + e_{22}\mu_{22})P(k_1, e_{11}, e_{12}, 0, e_{21}, e_{22}) \\
= \; & (e_{11} + 1)\theta_1\mu_{11}P(k_1, e_{11} + 1, e_{12} - 1, 0, e_{21}, e_{22}) \\
& + (e_{11} + 1)(1 - \theta_1)\mu_{11}P(k_1 - 1, e_{11} + 1, e_{12}, 0, e_{21}, e_{22}) \\
& + (e_{12} + 1)\mu_{12}P(k_1 - 1, e_{11}, e_{12} + 1, 0, e_{21}, e_{22}) \\
& + (e_{21} + 1)\theta_2\mu_{21}P(k_1, e_{11}, e_{12}, 0, e_{21} + 1, e_{22} - 1) \\
& + e_{21}(1 - \theta_2)\mu_{21}P(k_1 + 1, e_{11} - 1, e_{12}, 0, e_{21}, e_{22}) \\
& + (e_{22} + 1)\mu_{22}P(k_1 + 1, e_{11} - 1, e_{12}, 0, e_{21} - 1, e_{22} + 1) \qquad (5)
\end{aligned}$$

For $0 < k_1 = K_1 - c_1, k_2 = 0$ with $e_{11} + e_{12} = min(K_1 - k_1, c_1); e_{21} + e_{22} = c_2$:

$$\begin{aligned}
& (e_{11}\mu_{11} + e_{12}\mu_{12} + e_{21}\mu_{21} + e_{22}\mu_{22})P(K_1 - c_1, e_{11}, e_{12}, 0, e_{21}, e_{22}) \\
= \; & (e_{11} + 1)\theta_1\mu_{11}P(K_1 - c_1, e_{11} + 1, e_{12} - 1, 0, e_{21}, e_{22}) \\
& + e_{11}(1 - \theta_1)\mu_{11}P(K_1 - c_1 - 1, e_{11}, e_{12}, 0, e_{21}, e_{22}) \\
& + (e_{12} + 1)\mu_{21}P(K_1 - c_1 - 1, e_{11} - 1, e_{12} + 1, 0, e_{21}, e_{22}) \\
& + (e_{21} + 1)\theta_2\mu_{21}P(K_1 - c_1, e_{11}, e_{12}, 0, e_{21} + 1, e_{22} - 1) \\
& + e_{21}(1 - \theta_2)\mu_{21}P(K_1 - c_1 + 1, e_{11} - 1, e_{12}, 0, e_{21}, e_{22}) \\
& + (e_{22} + 1)\mu_{22}P(K_1 - c_1 + 1, e_{11} - 1, e_{12}, 0, e_{21} - 1, e_{22} + 1) \qquad (6)
\end{aligned}$$

For $0 < k_1 < K_1 - c_1, k_2 = 0$ with $e_{11} + e_{12} = min(K_1 - k_1, c_1); e_{21} + e_{22} = c_2$:

$$\begin{aligned}
& (e_{11}\mu_{11} + e_{12}\mu_{12} + e_{21}\mu_{21} + e_{22}\mu_{22})P(k_1, e_{11}, e_{12}, 0, e_{21}, e_{22}) \\
= \; & (e_{11}+1)\theta_1\mu_{11}P(k_1, e_{11}+1, e_{12}-1, 0, e_{21}, e_{22}) \\
& +e_{11}(1-\theta_1)\mu_{11}P(k_1-1, e_{11}, e_{12}, 0, e_{21}, e_{22}) \\
& +(e_{12}+1)\mu_{12}P(k_1-1, e_{11}-1, e_{12}+1, 0, e_{21}, e_{22}) \\
& +(e_{21}+1)\theta_2\mu_{21}P(k_1, e_{11}, e_{12}, 0, e_{21}+1, e_{22}-1) \\
& +e_{21}(1-\theta_2)\mu_{21}P(k_1+1, e_{11}, e_{12}, 0, e_{21}, e_{22}) \\
& +(e_{22}+1)\mu_{22}P(k_1+1, e_{11}, e_{12}, 0, e_{21}-1, e_{22}+1)
\end{aligned} \tag{7}$$

For $k_1 = 0, k_2 = 0$ with $e_{11} + e_{12} = c_1; e_{21} + e_{22} = c_2$:

$$\begin{aligned}
& (e_{11}\mu_{11} + e_{12}\mu_{12} + e_{21}\mu_{21} + e_{22}\mu_{22})P(0, e_{11}, e_{12}, 0, e_{21}, e_{22}) \\
= \; & (e_{11}+1)\theta_1\mu_{11}P(0, e_{11}+1, e_{12}-1, 0, e_{21}, e_{22}) \\
& + \begin{cases} e_{11}(1-\theta_1)\mu_{11}P(0, e_{11}, e_{12}, 1, e_{21}-1, e_{22}) & \text{if } c_2 = K_2 \\ e_{11}(1-\theta_1)\mu_{11}P(0, e_{11}, e_{12}, 1, e_{21}, e_{22}) & \text{if } c_2 < K_2 \end{cases} \\
& + \begin{cases} (e_{12}+1)\mu_{12}P(0, e_{11}-1, e_{12}+1, 1, e_{21}-1, e_{22}) & \text{if } c_2 = K_2 \\ (e_{12}+1)\mu_{12}P(0, e_{11}-1, e_{12}+1, 1, e_{21}, e_{22}) & \text{if} c_2 < K_2 \end{cases} \\
& +(e_{21}+1)\theta_2\mu_{21}P(0, e_{11}, e_{12}, 0, e_{21}+1, e_{22}-1) \\
& + \begin{cases} e_{21}(1-\theta_2)\mu_{21}P(1, e_{11}-1, e_{12}, 0, e_{21}, e_{22}) & \text{if } c_1 = K_1 \\ e_{21}(1-\theta_2)\mu_{21}P(1, e_{11}, e_{12}, 0, e_{21}, e_{22}) & \text{if } c_1 < K_1 \end{cases} \\
& + \begin{cases} (e_{22}+1)\mu_{22}P(1, e_{11}-1, e_{12}, 0, e_{21}-1, e_{22}+1) & \text{if } c_1 = K_1 \\ (e_{22}+1)\mu_{22}P(1, e_{11}, e_{12}, 0, e_{21}-1, e_{22}+1) & \text{if } c_1 < K_1 \end{cases}
\end{aligned} \tag{8}$$

For $k_1 = 0, K_2 - c_2 < k_2 \leq K_2$ with $e_{21} + e_{22} = min(K_2 - k_2, c_2); e_{11} + e_{12} = c_1$:

$$\begin{aligned}
& (e_{11}\mu_{11} + e_{12}\mu_{12} + e_{21}\mu_{21} + e_{22}\mu_{22})P(0, e_{11}, e_{12}, k_2, e_{21}, e_{22}) \\
= \; & (e_{11}+1)\theta_1\mu_{11}P(0, e_{11}+1, e_{12}-1, k_2, e_{21}, e_{22}) \\
& +e_{11}(1-\theta_1)\mu_{11}P(0, e_{11}, e_{12}, k_2+1, e_{21}-1, e_{22}) \\
& +(e_{12}+1)\mu_{12}P(0, e_{11}-1, e_{12}+1, k_2+1, e_{21}-1, e_{22}) \\
& +(e_{21}+1)\theta_2\mu_{21}P(0, e_{11}, e_{12}, k_2, e_{21}+1, e_{22}-1) \\
& +(e_{21}+1)(1-\theta_2)\mu_{21}P(0, e_{11}, e_{12}, k_2-1, e_{21}+1, e_{22}) \\
& +(e_{22}+1)\mu_{22}P(0, e_{11}, e_{12}, k_2-1, e_{21}, e_{22}+1)
\end{aligned} \tag{9}$$

For $k_1 = 0, 0 < k_2 = K_2 - c_2$ with $e_{21} + e_{22} = min(K_2 - k_2, c_2); e_{11} + e_{12} = c_1$:

$$\begin{aligned}
& (e_{11}\mu_{11} + e_{12}\mu_{12} + e_{21}\mu_{21} + e_{22}\mu_{22})P(0, e_{11}, e_{12}, K_2 - c_2, e_{21}, e_{22}) \\
= \; & (e_{11}+1)\theta_1\mu_{11}P(0, e_{11}+1, e_{12}-1, K_2 - c_2, e_{21}, e_{22}) \\
& + e_{11}(1-\theta_1)\mu_{11}P(0, e_{11}, e_{12}, K_2 - c_2 + 1, e_{21} - 1, e_{22}) \\
& + (e_{12}+1)\mu_{12}P(0, e_{11}-1, e_{12}+1, K_2 - c_2 + 1, e_{21} - 1, e_{22}) \\
& + (e_{21}+1)\theta_2\mu_{21}P(0, e_{11}, e_{12}, K_2 - c_2, e_{21} + 1, e_{22} - 1) \\
& + (e_{21}+1)(1-\theta_2)\mu_{21}P(0, e_{11}, e_{12}, K_2 - c_2 - 1, e_{21}, e_{22}) \\
& + (e_{22}+1)\mu_{22}P(0, e_{11}, e_{12}, K_2 - c_2 - 1, e_{21} - 1, e_{22} + 1)
\end{aligned} \tag{10}$$

For $k_1 = 0, 0 < k_2 < K_2 - c_2$ with $e_{21} + e_{22} = min(K_2 - k_2, c_2); e_{11} + e_{12} = c_1$:

$$\begin{aligned}
& (e_{11}\mu_{11} + e_{12}\mu_{12} + e_{21}\mu_{21} + e_{22}\mu_{22})P(0, e_{11}, e_{12}, k_2, e_{21}, e_{22}) \\
= \; & (e_{11}+1)\theta_1\mu_{11}P(0, e_{11}+1, e_{12}-1, k_2, e_{21}, e_{22}) \\
& + e_{11}(1-\theta_1)\mu_{11}P(0, e_{11}, e_{12}, k_2 + 1, e_{21}, e_{22}) \\
& + (e_{12}+1)\mu_{12}P(0, e_{11}-1, e_{12}+1, k_2 + 1, e_{21}, e_{22}) \\
& + (e_{21}+1)\theta_2\mu_{21}P(0, e_{11}, e_{12}, k_2, e_{21} + 1, e_{22} - 1) \\
& + (e_{21}+1)(1-\theta_2)\mu_{21}P(0, e_{11}, e_{12}, k_2 - 1, e_{21}, e_{22}) \\
& + (e_{22}+1)\mu_{22}P(0, e_{11}, e_{12}, k_2 - 1, e_{21} - 1, e_{22} + 1)
\end{aligned} \tag{11}$$

Finally, the normalization equation implies

$$\sum_{s \in \mathbb{S}} P(k_1, e_{11}, e_{12}, k_2, e_{21}, e_{22}) = 1 \tag{12}$$

Solving the system of equations (Equation 5 to Equation 12), the steady state probabilities $P(k_1,\ e_{11}, e_{12},\ k_2, e_{21}, e_{21})$ can be derived for each state in $\mathbb{S}$. Using the steady state probabilities, expressions for throughput and mean queue lengths are written. The throughput is given by:

$$\begin{aligned}
\lambda_D \; = \; & \sum_{s \in \mathbb{S}_1} P(k_1, e_{11}, e_{12}, k_2, e_{21}, e_{22}) \left[e_{21}(1-\theta_2)\mu_{21} + e_{22}\mu_{22}\right] \\
+ \; & \sum_{s \in \mathbb{S}_2} P(0, e_{11}, e_{12}, k_2, e_{21}, e_{22}) \left[e_{11}(1-\theta_1)\mu_{11} + e_{12}\mu_{12}\right]
\end{aligned} \tag{13}$$

The average queue length in input buffer B_1 is given by:

$$E(L_1) = \sum_{s \in \mathbb{S}_1} k_1 P(k_1, e_{11}, e_{12}, 0, e_{21}, e_{22}) \tag{14}$$

The average queue length in input buffer B_2 is given by:

$$E\left(L_2\right) = \sum_{s \in \mathbb{S}_2} k_2 P(0, e_{11}, e_{12}, k_2, e_{21}, e_{22}) \tag{15}$$

Note that the system of equations (Equation 5 to Equation 12) permit the exact analysis of a wide class of fork/join stations. For instance, by setting one or both of the c_i's equal to 1 (for $i = 1, 2$), the results for a fork/join station with inputs from a single server station can be obtained. Additionally, by suitably choosing parameters for the 2-phase Coxian distribution, systems with exponentially distributed service times could be analyzed. Also, if $K_i = c_i$ for $i = 1, 2$, the system could be used to analyze a two-stage tandem line with multiple servers and zero buffers.

5 Effect of Variability on Performance Measures

This exact analysis described in Section 4 is to used analyze the the effect of variability on performance measures. The discussion is structured as follows. Section 5.1 discusses systems wherein both the inputs to the fork/join station are from stations with multiple servers. In the remainder of the paper, the notation MM is used to denote this configuration. Subsequently, Section 5.2 presents the analysis of systems wherein one of the inputs is from a station with multiple servers, while the other is from a station with a single server. The notation MS is used to denote this configuration. Section 5.3 presents the analysis of systems wherein both the inputs are from stations with a single server. The notation SS is used to denote this configuration. Note that exact analysis of each of these systems (MM, MS and SS) can be carried out using the equations derived in the previous section. Section 5.4 compares the performance of all three systems with respect to the effect of variability on throughput and mean queue lengths.

5.1 Variability Effects in MM Systems

Using the equations presented in the previous section numerical results are obtained to analyze the effect of variability on the performance of MM systems. Three experiments are conducted. In all three experiments, the number of servers, c_i, at station $i = 1, 2$ are

kept equal to the total population size, K_i. The first experiment, $MM(i)$ corresponds to a balanced system, wherein station capacities defined by $K_1\mu_1$ (at station 1) and $K_2\mu_2$ (at station 2) are set equal to one, i.e. $K_1\mu_1 = K_2\mu_2 = 1$. The population size K_i is varied to take values of $K_i = 2, 5, 10$ and 25 respectively, while maintaining $K_1 = K_2$. For each of these four settings, the SCV of service times, $c^2_{S,i}, i = 1, 2$ is varied to take values of $c^2_{S,i} = 0.5, 1.0, 1.5, 2.0$ and 2.5 while maintaining $c^2_{S,1} = c^2_{S,2}$. These combinations yield a total of 20 settings as part of experiment $MM(i)$. Note that although the values of K_i are varied in these 20 settings, the station service capacities are always equal to 1. For instance, when $K_1 = K_2 = 10$, then $\mu_1 = \mu_2 = 0.1$, so that $K_1\mu_1 = K_2\mu_2 = 1$. Further, when $K_1 = K_2 = 25$, then $\mu_1 = \mu_2 = 0.04$. Maintaining station service rates equal, permits a fair comparison between the 20 settings. The objective of the second experiment, $MM(ii)$ is to investigate the effect of variations in SCV of only one of the inputs on the key performance measures. Again, station capacities are set equal to one, i.e. $K_1\mu_1 = K_2\mu_2 = 1$, and population size K_i are varied to take values of $K_i = 2, 5, 10$ and 25 respectively, while maintaining $K_1 = K_2$. However, unlike experiment $MM(i)$ only one of the SCVs, $c^2_{S,2}$ is varied to take values 0.5, 1.0, 1.5, 2.0 and 2.5 while maintaining $c^2_{S,1} = 1$. The goal of the third experiment, $MM(iii)$ is to investigate the effect of variations in SCV when station capacities are imbalanced. In this experiment, $K_1\mu_1 = 1$ while $K_2\mu_2 = 1.25$, and population size K_i are varied to take values of $K_i = 2, 5, 10$ and 25 respectively, while maintaining $K_1 = K_2$. As in experiment $MM(ii)$, $c^2_{S,2}$ is varied to take values 0.5, 1.0, 1.5, 2.0 and 2.5 while maintaining $c^2_{S,1} = 1$. The results of experiments $MM(i)$, $MM(ii)$ and $MM(iii)$ are reported in Figure 2. The figure plots the throughput, λ_D, and mean queue lengths, $E(L_1)$ and $E(L_2)$, for each experiment. From the figure the following observations can be made:

(i) The throughput, λ_D, is non-increasing with SCV ($c^2_{S,i}, i = 1, 2$), and non-decreasing with K_i. Further, the effect of SCVs on throughput appears to diminish with increase in K_i. The figures also indicate that when system capacities are balanced, the mean queue lengths $E(L_1)$ and $E(L_2)$ are non-decreasing with SCV and K_i. As in the case of throughput, the relative effect of SCVs on mean queue lengths also diminish with increase in K_i.

(ii) Experiment $MM(i)$ and $MM(ii)$ suggest that the effect of SCV on performance measures is relatively more when they are varied simultaneously for the inputs to both buffers of the fork/join station.

(iii) As expected, the throughput of the imbalanced system in experiment $MM(iii)$ is comparatively higher than that of the corresponding balanced system in experiment $MM(ii)$. The throughput of the fork/join station approaches the service rate of the

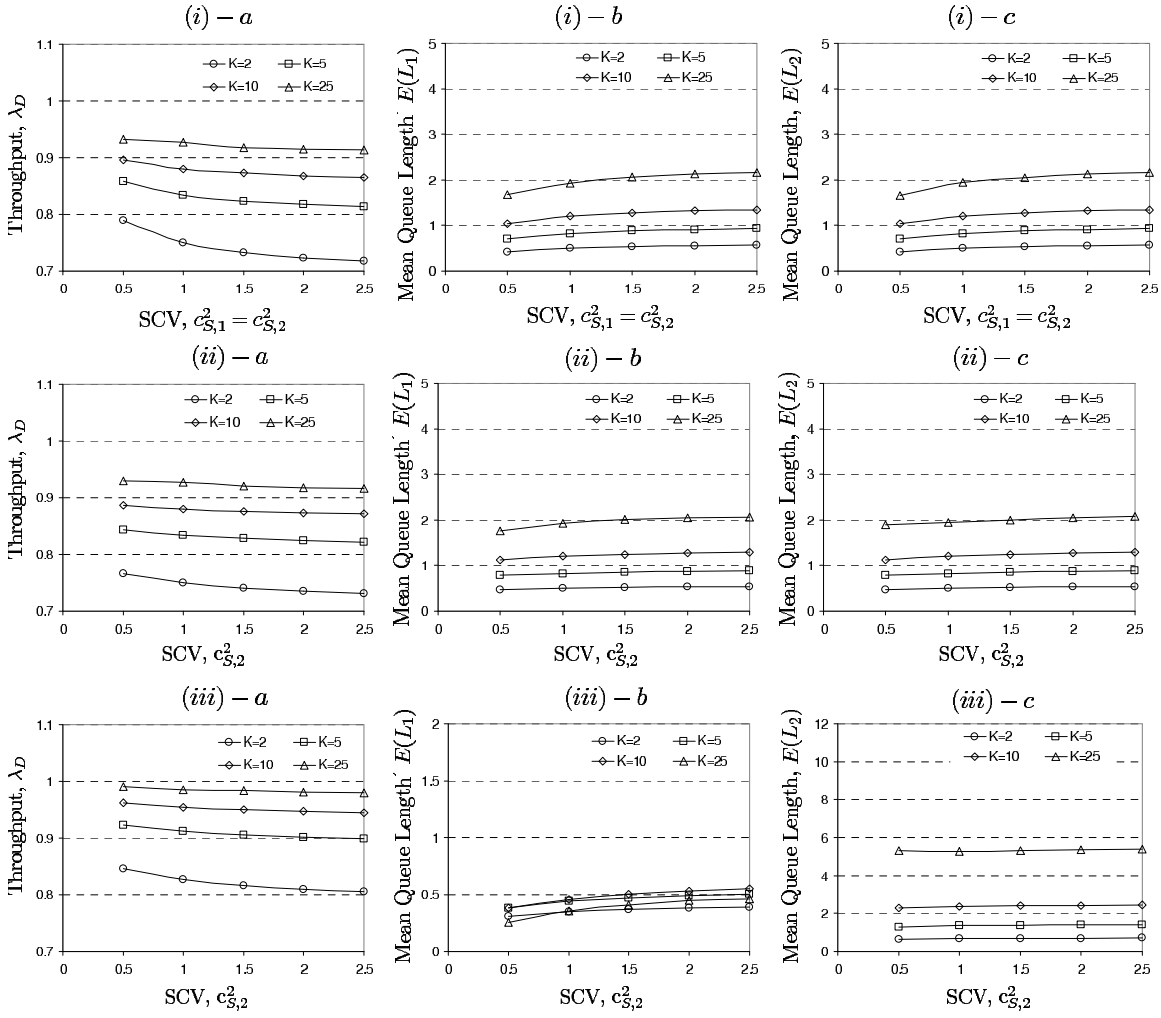


Figure 2: Impact of variability in MM systems (Cases $(i) - a$, $(i) - b$ and $(i) - c$ correspond to Experiment $MM(i)$, cases $(ii) - a$, $(ii) - b$ and $(ii) - c$ correspond to Experiment $MM(ii)$, and cases $(iii) - a$, $(iii) - b$ and $(iii) - c$ correspond to Experiment $MM(iii)$ respectively)

slowest station $\min(K_1\mu_1, K_2\mu_2)$ with increase in K_i values. The figures also suggest that capacity imbalance leads to unequal distribution of queue lengths, $E(L_1)$ and $E(L_2)$. Moreover, it appears that capacity imbalances dominate over influence of SCV variations on performance measures.

5.2 Variability Effects in MS Systems

Next, the results from numerical experiments conducted for MS systems are discussed. Four experiments are conducted. In all experiments, station 1 consists of multiple servers, while station 2 consists of a single server. Further, in all experiments, the number of servers, c_1, at station 1 is kept equal to the total population size, K_1. The first experiment, $MS(i)$

corresponds to a system with balanced station capacities, i.e. $K_1\mu_1 = \mu_2 = 1$. A total of 20 settings are considered. In these settings, the population size K_i is varied to take values of $K_i = 2, 5, 10$ and 25 respectively (while maintaining $K_1 = K_2$), and the SCV of service times, $c^2_{S,i}, i = 1, 2$ are varied to take values of $c^2_{S,i} = 0.5, 1.0, 1.5, 2.0$ and 2.5 (while maintaining $c^2_{S,1} = c^2_{S,2}$). The objective of the second experiment, $MS(ii)$ is to investigate the effect of variations in SCV of only one of the inputs on the key performance measures. Again, station capacities are set equal to one, i.e. $K_1\mu_1 = \mu_2 = 1$, and population size K_i is varied to take values of between 2 and 25, and $c^2_{S,1}$ is varied to take values between 0.5 and 2.5 while maintaining $c^2_{S,2} = 1$. The goal of the third and fourth experiments $MS(iii)$ and $MS(iv)$ is to investigate the effect of variations in SCV when station capacities were imbalanced. In experiment $MS(iii)$, the single server station, station 2, has a higher capacity of 1.25, while the multi-server station, station 1 has a capacity of 1. In experiment $MS(iv)$, station capacities are reversed with the multi-server station, station 1 having a capacity equal to 1.25. In both $MS(iii)$ and $MS(iv)$, the population size K_i is varied to take values of $K_i = 2, 5, 10$ and 25 respectively, while maintaining $K_1 = K_2$. Also $c^2_{S,1}$ is varied to take values 0.5, 1.0, 1.5, 2.0 and 2.5 while maintaining $c^2_{S,2} = 1$. The results of experiments $MS(i)$ and $MS(ii)$ are reported in Figure 3 while the results of $MS(iii)$ and $MS(iv)$ are reported in Figure 4. The figure plots the throughput, λ_D, and mean queue lengths, $E(L_1)$ and $E(L_2)$, for each experiment. From the figure the following observations can be made:

(i) As with the MM systems, the throughput, λ_D, in an MS system is a non-increasing with SCV, and non-decreasing with K_i. Further, the effect of SCVs on throughput appears to diminish with increase in K_i. In balanced systems, the effect of SCV on throughput appears to be more than that observed for MM systems in the previous section. However, in systems with imbalances in station capacities, the effect of SCVs seems to be significant only for low values of K_i.

(ii) The mean queue length $E(L_1)$ is non-decreasing with SCV. However, depending on the system configuration, the mean queue length $E(L_2)$ could either increase or decrease with increase in SCV. Experiments $MS(i) - MS(ii)$ suggest that when station capacities are balanced, $E(L_2)$ is non-increasing in SCV. A similar behavior is observed in experiment $MS(iii)$ where station 2 has a larger station capacity. However, in experiment $MS(iv)$, when station 2 has a smaller station capacity, $E(L_2)$ is non-decreasing in SCV.

(iii) Unlike the MM system, even when station capacities are balanced, as in experiment $MS(i)$, the mean queue lengths $E(L_1)$ and $E(L_2)$ need not be equal. This is because, in an MS system, even if capacities at stations 1 and 2 are equal, the service rates at

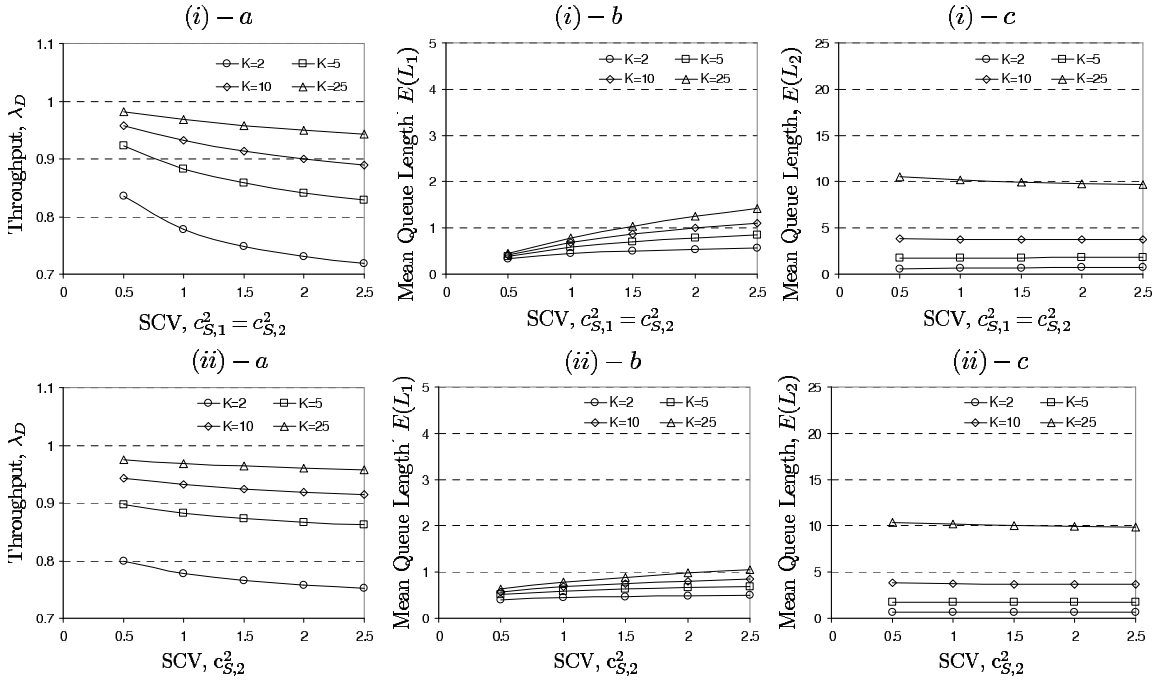


Figure 3: Impact of variability in MS systems with balanced capacities (Cases $(i)-a$, $(i)-b$ and $(i)-c$ correspond to Experiment $MS(i)$ and cases $(ii)-a$, $(ii)-b$ and $(ii)-c$ correspond to Experiment $MS(ii)$ respectively)

each station could be different. Since, station 2 is a single server station, the service rate is equal to 1 whenever the station is not idle. However, in the case of station 1, the service rate of a station is equal to 1 only when all the servers at the station are busy. Consequently, $E(L_1) \leq E(L_2)$ even when station capacities are balanced.

(iv) As with the MM systems, the throughput of the imbalanced system in experiment $MS(iii)$ and $MS(iv)$ are comparatively higher than that of the corresponding balanced system in experiment $MS(ii)$. The throughput of the fork/join station tends to the service rate of the slowest station $\min(K_1\mu_1, \mu_2)$ with increase in K_i values. While capacity imbalance leads to unequal distribution of queue lengths, the effect of SCVs appears to to be relatively less in unbalanced systems. As in the MM systems, it appears that capacity imbalances dominate over influence of SCV variations on performance measures.

5.3 Variability Effects in SS Systems

Next, the results from numerical experiments conducted for SS systems are discussed. Three experiments are conducted. In all experiments, both station 1 and station 2 consist of a

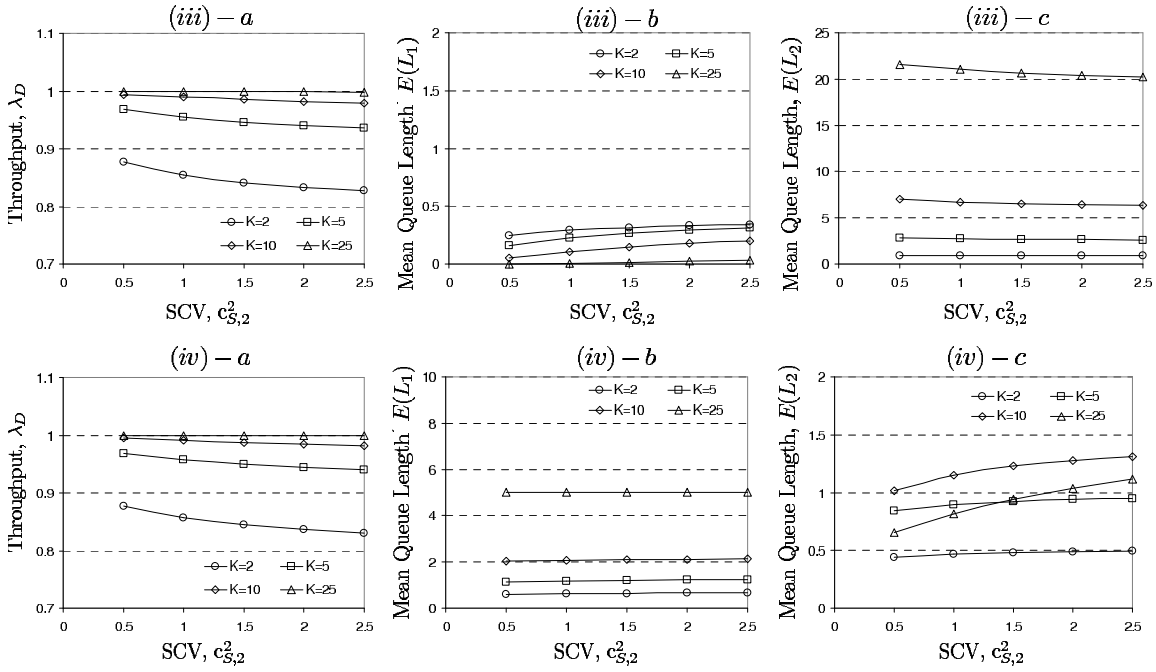


Figure 4: Impact of variability in MS systems with unbalanced capacities (Cases $(iii)-a$, $(iii)-b$ and $(iii)-c$ correspond to Experiment $MS(iii)$ and cases $(iv)-a$, $(iv)-b$ and $(iv)-c$ correspond to Experiment $MS(iv)$ respectively)

single server. The first experiment, $SS(i)$ corresponds to a system with balanced station capacities, i.e. $\mu_1 = \mu_2 = 1$. Again, 20 settings are considered wherein the population size K_i is varied to take values of $K_i = 2, 5, 10$ and 25 respectively (while maintaining $K_1 = K_2$), and the SCV of service times, $c^2_{S,i}, i = 1, 2$ are varied to take values of $c^2_{S,i} = 0.5, 1.0, 1.5, 2.0$ and 2.5 (while maintaining $c^2_{S,1} = c^2_{S,2}$). The objective of the second experiment, $SS(ii)$ is to investigate the effect of variations in SCV of only one of the inputs on the key performance measures. Again, station capacities are set equal to one, i.e. $\mu_1 = \mu_2 = 1$, and population size K_i is varied to take values between 2 and 25, and $c^2_{S,2}$ is varied to take values between 0.5 and 2.5 while maintaining $c^2_{S,1} = 1$. The goal of third experiment, $SS(iii)$ is to investigate the effect of variations in SCV of when station capacities are unbalanced. Consequently, station 2 had a higher capacity of 1.25, while station 1 had a capacity of 1. Again, the population size K_i is varied to take values of $K_i = 2, 5, 10$ and 25 respectively, while maintaining $K_1 = K_2$. Also $c^2_{S,2}$ is varied to take values 0.5, 1.0, 1.5, 2.0 and 2.5 while maintaining $c^2_{S,1} = 1$. The results of experiments $SS(i) - SS(iii)$ are reported in Figure 5. From the figure the following observations are made:

(i) As with the MM and MS systems, the throughput, λ_D, in an SS system is non-increasing with SCV, and non-decreasing with K_i. As in the other systems, the effect of SCVs

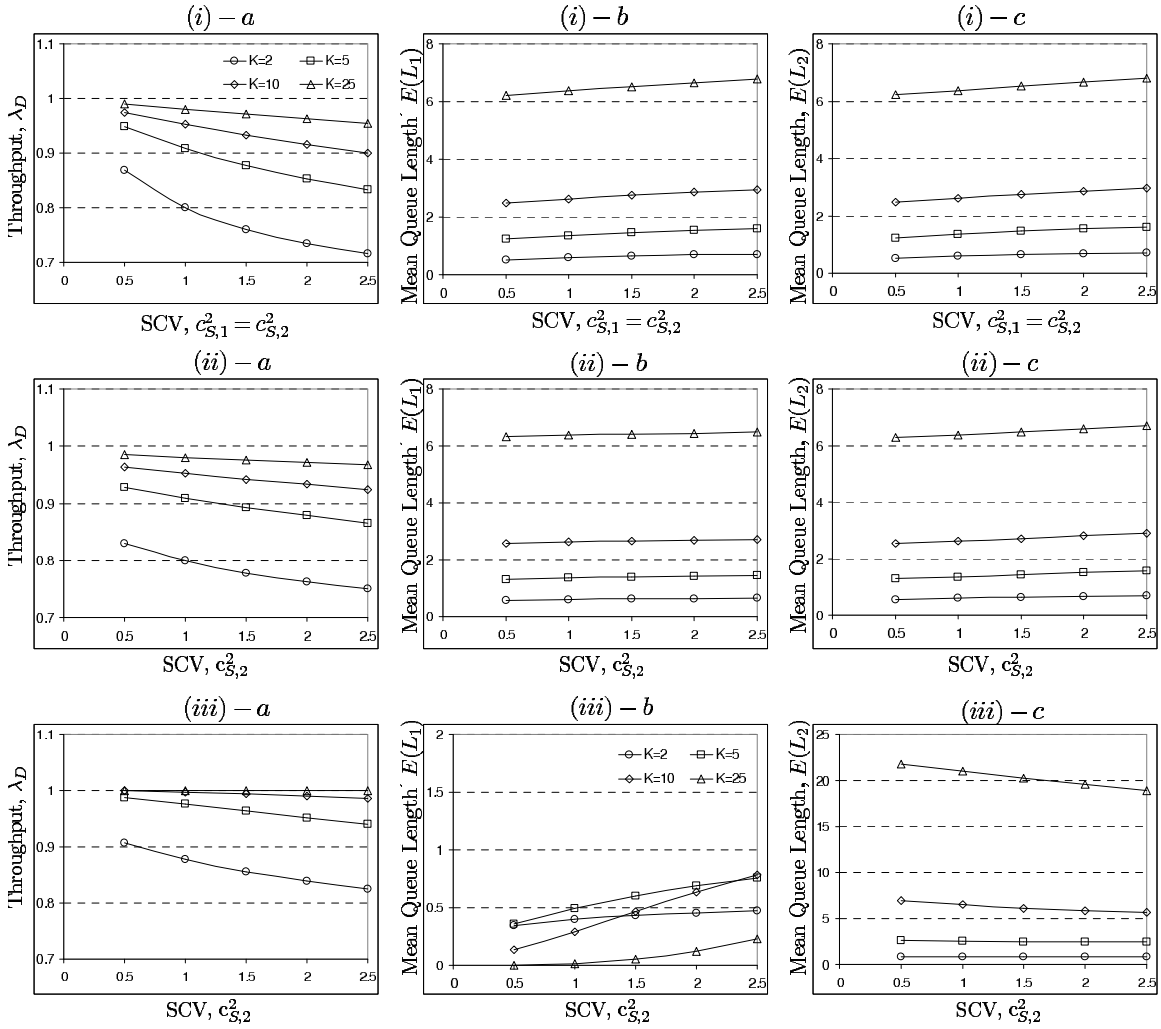


Figure 5: Impact of variability in SS systems (Cases $(i) - a$, $(i) - b$ and $(i) - c$ correspond to Experiment $SS(i)$, cases $(ii) - a$, $(ii) - b$ and $(ii) - c$ correspond to Experiment $SS(ii)$, and cases $(iii) - a$, $(iii) - b$ and $(iii) - c$ correspond to Experiment $SS(iii)$ respectively)

on throughput appears to diminish with increase in K_i. However, in balanced systems, the effect of SCV on throughput appears to be more than that observed for MM or MS systems.

(ii) The mean queue length $E(L_1)$ is non-decreasing with SCV. However, depending on the system configuration, the mean queue length $E(L_2)$ could either increase or decrease with increase in SCV. Experiments $SS(i) - SS(ii)$ suggest that when station capacities are balanced, $E(L_2)$ is non-decreasing in SCV. However, in experiment $SS(iii)$, when station 2 has a higher station capacity, $E(L_2)$ is non-increasing in SCV.

(iii) As with the MM and MS systems, the throughput of the imbalanced system in experi-

ment $SS(iii)$ are comparatively higher than that of the corresponding balanced system in experiment $SS(ii)$. The throughput of the fork/join station tends approaches the service rate of the slowest station $\min(\mu_1, \mu_2)$ with increase in K_i values. While capacity imbalance leads to unequal distribution of queue lengths, the effect of SCVs appears to be relatively less in unbalanced systems.

5.4 Performance Comparison of Fork/Join Systems

This section provides a brief comparison of the performance of MM, MS, and SS systems. Figure 6 plots the throughput, λ_D, and mean queue lengths, $E(L_1)$ and $E(L_2)$, for MM, MS and SS systems in two settings. In both settings, station capacities at station 1 and 2 are set equal to one and the SCV of service times, $c^2_{S,i}, i = 1, 2$ are varied to take values of $c^2_{S,i} = 0.5, 1.0, 1.5, 2.0$ and 2.5 (while maintaining $c^2_{S,1} = c^2_{S,2}$). In the first set of experiments, the population size K_i is set equal to two, i.e. $K_1 = K_2 = 2$, while in the second set of experiments, the population size K_i is set equal to ten, i.e. $K_1 = K_2 = 10$. In the discussion below, a superscript of MM, MS and SS is used to denote the performance measure of the respective systems. From the figures, the following observations are made.

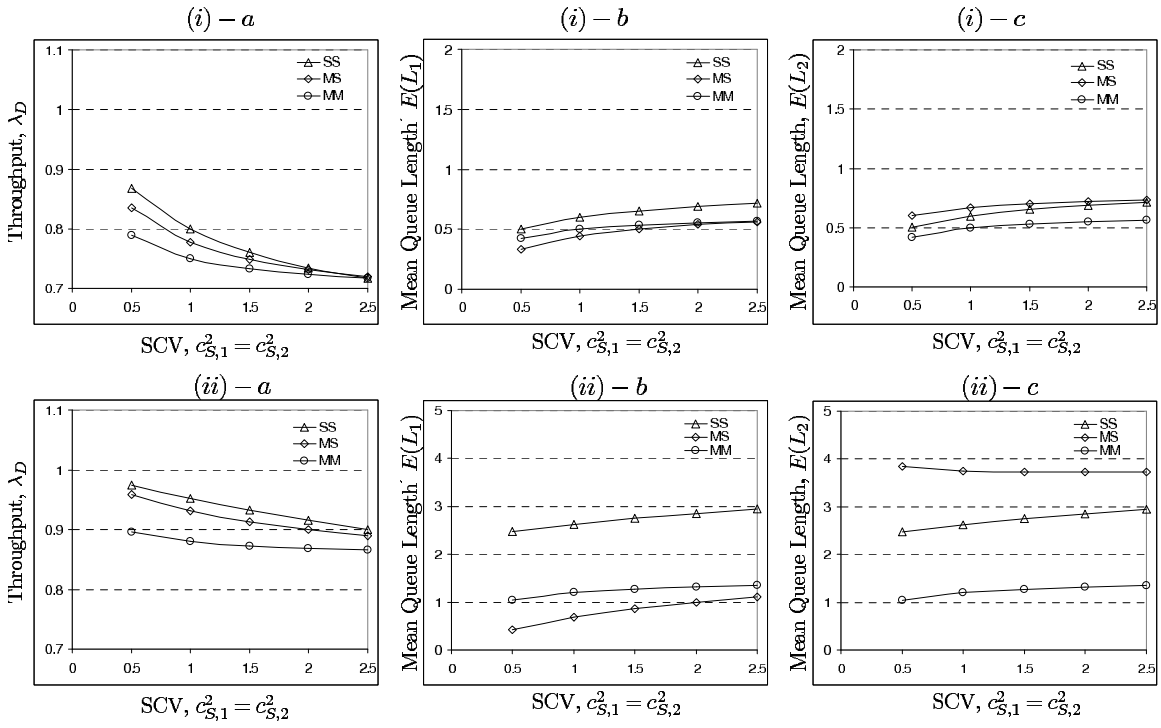


Figure 6: Performance comparison of MM, MS and SS systems (Cases $(i) - a$, $(i) - b$ and $(i) - c$ correspond to $K_1 = K_2 = 2$ and cases $(ii) - a$, $(ii) - b$ and $(ii) - c$ correspond to $K_1 = K_2 = 10$)

(i) $\lambda_D^{MM} \leq \lambda_D^{MS} \leq \lambda_D^{SS}$: When station capacities are the same, for a given population size, the throughput of the SS system is the highest while that of the MM system is the lowest. This is because, the service rates at each station in the SS system is equal to 1 when the station is not idle. In contrast, for the MM system, the service rate of a station is equal to 1 only when all the servers at the station are busy.

(ii) In all systems, MM, MS and SS, it appears that the effect of SCV on throughput, λ_D, is more significant than its effect on mean queue lengths $E(L_1)$ and $E(L_2)$. Moreover, it appears that the effect of SCV on the throughput of MM systems is less when compared to the effect of SCV on the throughput of SS systems. In all systems the effect of SCV on performance measures decreases with increase in K_i.

(iii) $E(L_i^{SS}) \geq E(L_i^{MM}), i = 1, 2$: When station capacities are the same, the mean queue length at buffers B_1 and B_2 is always higher for the SS system than for the MM system. Although station capacities at each station is equal to one in both MM and SS systems, the service rates at the stations need not be equal. In the SS system, the service rate of a station is equal to 1 when the single server at the station is not idle. However, in the case of the MM system, the service rate of a station is equal to 1 only when all the servers at a station are busy. This results in less built up of queues at buffers B_1 and B_2.

(iv) $E(L_1^{SS}) = E(L_2^{SS})$ and $E(L_1^{MM}) = E(L_2^{MM})$ but $E(L_1^{MS}) \leq E(L_2^{MS})$: For SS and MM systems, when station capacities, population limits, and SCVs are the same, the mean queue length at buffer B_1 is equal to the mean queue length at buffer B_2. For MS systems, even when station capacities are the same, the mean queue length at the buffer following multi-server station in an MS system (i.e. buffer B_1) is always less than the mean queue length at the buffer following single server station (i.e. buffer B_2). In the MS system, the imbalance in station rates, even when station capacities are balanced, leads to excess queues at buffer B_2.

(v) $E(L_1^{MS}) \leq E(L_1^{MM})$ and $E(L_2^{MS}) \geq E(L_2^{SS})$: The mean queue length at the buffer following multi-server station in an MS system (i.e. buffer B_1) is always less than the mean queue length at the corresponding buffer in an MM system having identical station capacities, population limits, and SCVs. Similarly, the mean queue length at the buffer following single server station in an MS system (i.e. buffer B_2) is always higher than the mean queue length at the corresponding buffer in an SS system having identical station capacities, population limits, and SCVs. In the MS system defined above, station 2 operates at a rate of one when it is not idle. However, station 1

operates at a rate of one only when all servers at the station are busy. At all other times, the station rate is strictly less than one.

6 Approximations for Performance Measures

The performance analysis and numerical comparisons discussed in the above section indicate that the influence of SCV on the throughput and mean queue lengths could be relatively less in many settings for MM and MS systems. This suggests that in these settings, if exact estimates were not essential, and instead, reasonably accurate estimates of performance measures would be adequate, then, simpler approximations could be used. In particular, performance estimates of a system with exponential inputs could be used as approximations. Such systems are relatively simple to analyze. This section first describes the analysis of a system with exponential inputs and then investigates the accuracy of the approximations developed based on that analysis.

As before, let $N_1(t)$ and $N_2(t)$denote the number of units in buffers B_1 and B_2 respectively at time t. Then, for the case of exponential inputs, the state of the system is characterized by $(k_1, k_2) = [N_1(t) = k_1, N_2(t) = k_2]$, $t \geq 0$. Clearly, (k_1, k_2) is a continuous time Markov chain defined on the state space $[(K_1, 0), (K_1 - 1, 0) \ldots, (1, 0), (0, 0), (0, 1), \ldots, (0, K_2 - 1), (0, K_2)]$. Therefore the state transition rates for the continuous time Markov chain representing the queue length process are illustrated in Figure 7. It can be shown that this Markov chain is positive recurrent. Therefore, the steady state probability of state (k_1, k_2) given by $P(k_1, k_2)$ can be obtained by solving the set of balance equations.
In terms of these probabilities, expressions for throughput and the mean queue lengths at buffers B_1 and B_2 are given by:

$$\lambda_D = \left[K_1 \mu_1 \sum_{k_2=1}^{K_2} P(0, k_2) + K_2 \mu_2 \sum_{k_1=1}^{K_1} P(k_1, 0) \right] \tag{16}$$

and :

$$E(L_1) = \sum_{k_1=1}^{K_1} k_1 P(k_1, 0) \quad \text{and} \quad E(L_2) = \sum_{k_2=1}^{K_2} k_2 P(0, k_2) \tag{17}$$

Clearly, the system with exponential inputs only requires the solution of a Markov chain with $K_1 + K_2 + 1$ states. This computational advantage might be attractive if the analysis of the fork/join station is being carried out in the context of larger closed queuing networks with fork/join stations. To investigate the regions in the design space where such an approx-

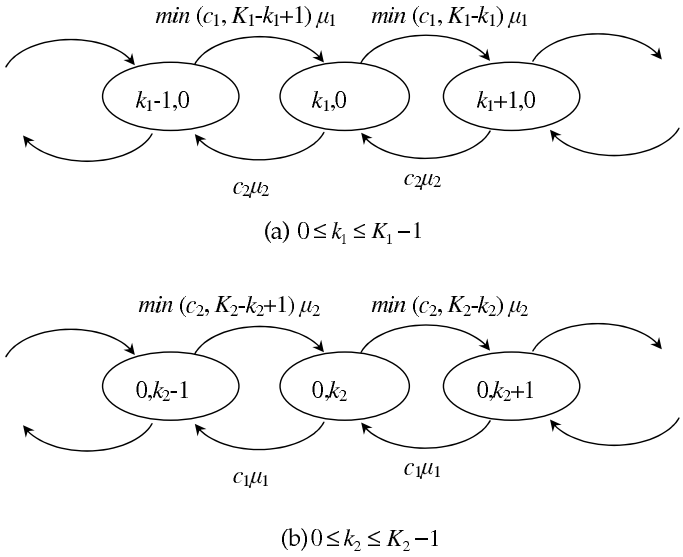


Figure 7: Rate balance for system with exponential inputs

imation would yield reasonably accurate performance estimates a full factorial experiment is carried out, for the MM and the MS systems. In each experiment, the exact values of the throughput λ_D^C, and mean queue lengths $E(L_1)^C$ and $E(L_2)^C$ are recorded using the exact analysis for Coxian inputs. These results are compared to the performance estimates (λ_D^E, $E(L_1)^E$, and $E(L_2)^E$) obtained for the corresponding system with exponential inputs (i.e., a system where the SCVs $c_{S,1}^2$ and $c_{S,2}^2$ are set equal to 1). The percentage error (δ, ϵ_1, and ϵ_2) between these estimates is used as a measure to determine the efficiency of approximations based on exponential inputs. These measures are defined by the following expressions:

$$\begin{aligned} \delta &= \frac{\left|\lambda_D^C - \lambda_D^E\right|}{\lambda_D^C}\% \\ \epsilon_i &= \frac{\left|E(L_i)^C - E(L_i)^E\right|}{K_i}\% \qquad i = 1, 2 \end{aligned} \tag{18}$$

The error in the queue length is computed as a percentage of K_i to avoid the potential problems that might arise when the mean queue length itself is small. The experiment design and results for MM and MS systems are summarized in the sections below.

6.1 Exponential System Approximation for MM systems

Table 2 shows the input parameters ranges used in the full factorial experiment. In total 900 experiments are conducted. As seen from the table, the experiment design considered 4

different capacity combinations ($K_1\mu_1$ and $K_2\mu_2$), 25 different SCV combinations ($c^2_{S,1}$ and $c^2_{S,2}$), and 9 different combinations of the population constraint (K_1 and K_2). In all the experiments, the population constraint is set equal to the number of servers at each station, i.e. $K_i = c_i$ for $i = 1, 2$.

$K_1\mu_1$	K_1	$c^2_{S,1}$	$K_2\mu_2$	K_2	$c^2_{S,2}$
1.25	2	0.5	1.25	2	0.75
1	5	1	1	5	1.25
	10	1.5		10	1.75
		2			2.25
		2.5			2.75

Table 2: Design of experiment for MM systems

Table 3 summarizes the results from the experiments. The table reports the average as well as the maximum values of the percentage errors. In addition to reporting the overall errors, the table also documents how these errors vary with station capacities, number of servers, population constraints, and SCVs. From the table, the following observations are made:

(i) From the overall percentage errors reported in the table, it appears that the exponential system provides reasonably good estimates of performance measures. For instance, the average errors over all the estimates of throughput, and mean queue lengths is less than 2 % with the maximum error being below 5%. In systems with imbalances in station capacities, it appears that the errors are marginally better. Also, the estimates in the mean queue length at the buffer following the station with higher capacity seems to be marginally more accurate.

(ii) With respect to the influence of population constraint and number of servers on the percentage errors, it is evident that the errors decrease with increase in K_i and c_i. For instance when $K_1 = K_2 = 10 = c_1 = c_2$, the average error in throughput and mean queue length estimates is less than 1%, with the maximum error being less than 2%. This implies that the effect of SCVs of the inputs diminishes as K_i and c_i increase, and the system behaves more like a system with exponential inputs under these conditions.

(iiii) The effect of SCV of inputs at low values of K_i and c_i can also be discerned from the table. It is evident that the errors (average and maximum) in estimates of throughput and mean queue lengths increase marginally as inputs have SCVs different from 1. However, even where the errors are marginally higher, they are never more than 5%.

	Average			Maximum		
	δ	ϵ_1	ϵ_2	δ	ϵ_1	ϵ_2
Overall	1.532	1.221	1.222	4.775	3.598	3.607

Variation with respect station capacities

	Average			Maximum		
$K_1\mu_1, K_2\mu_2$	δ	ϵ_1	ϵ_2	δ	ϵ_1	ϵ_2
1,1	1.618	1.299	1.299	4.775	3.418	3.418
1, 1.25	1.443	1.268	1.015	4.551	3.598	2.878
1.25, 1	1.451	1.019	1.274	4.562	2.885	3.607
1.25, 1.25	1.618	1.299	1.299	4.775	3.418	3.418

Variation with respect population constraints

	Average			Maximum		
$K_1 + K_2$	δ	ϵ_1	ϵ_2	δ	ϵ_1	ϵ_2
4	2.404	1.749	1.749	4.775	3.598	3.607
7	1.899	1.480	1.479	3.747	2.891	3.008
10	1.277	1.042	1.043	2.592	2.107	2.107
12	1.673	1.364	1.365	3.535	2.860	3.095
15	1.076	0.906	0.906	2.224	1.872	1.872
20	0.814	0.701	0.702	1.727	1.494	1.494

Variation with respect SCVs

	Average			Maximum		
$0.5(c_{S,1}^2 + c_{S,2}^2)$	δ	ϵ_1	ϵ_2	δ	ϵ_1	ϵ_2
0.625	1.835	1.524	1.514	3.309	2.808	2.567
0.875	0.628	0.516	0.512	1.684	1.512	1.308
1.125	0.532	0.433	0.433	1.696	1.357	1.539
1.375	1.002	0.804	0.805	2.691	2.195	2.409
1.625	1.430	1.140	1.141	3.535	2.860	3.095
1.875	1.895	1.504	1.505	3.468	2.808	2.958
2.125	2.287	1.809	1.810	3.952	2.993	3.007
2.375	2.591	2.044	2.046	4.401	3.335	3.324
2.625	2.839	2.234	2.235	4.775	3.598	3.607

Table 3: Summary of results for MM systems

To test whether the efficiency of the exponential approximation is sensitive to the choice of distribution, we evaluated the performance of a fork/join system where the inputs have lognormal distribution. Since exact analytical models are not available for these inputs, the estimates of throughput and mean queue lengths were obtained from a simulation model built in Arena (www.arenasimulation.com) for 72 scenarios. In particular, $K_1\mu_1$ was set

equal to 1 and $K_2\mu_2$ was permitted to take values from the set (1, 1.25). In these experiments, the parameters are chosen so that $c^2_{S,1}$ take values from the set (0.5, 1.5, and 2.5) while $c^2_{S,2}$, take values from the set (0.75, 1.75, and 2.75). In each case, the parameters of the lognormal distributions are chosen appropriately. In addition, K_1 and K_2 take values from the set (2, 10), and in all 72 settings, K_i is set equal to c_i. The results are summarized in Table 4 below. From the table it is evident that the average and maximum percentage errors are similar to those obtained for the case of Coxian inputs, suggesting that the efficiency of the exponential approximation is reasonable for different distributions of the inputs.

Average			Maximum		
δ	ϵ_1	ϵ_2	δ	ϵ_1	ϵ_2
1.270	1.100	0.970	4.980	4.080	3.920

Table 4: Performance of approximation for MM systems with lognormal inputs

The results from these experiments indicate that very efficient approximations for performance measures of an MM system can be obtained by analyzing the simpler system with exponential inputs. The performance estimates appear to be fairly robust across choice of distribution.

To identify specific zones where the error due to approximations increase in a relative sense, the percentage errors are plotted against $K_1 + K_2$ and $0.5(c^2_{S,1} + c^2_{S,2})$ in Figure 8 for throughput, λ_D, and mean queue lengths $E(L_1)$ and $E(L_2)$. It appears from these plots that in general the errors decrease, as K_i or c_i increase. The only region in the design space where the errors may be marginally high for practical decision making would be when $K_1 + K_2 \leq 6$ and $0.5(c^2_{S,1} + c^2_{S,2}) \geq 2.5$. However, in this region, the low values of $K_1 + K_2$ suggest that an exact analysis of a system with Coxian inputs might be possible without significant computation effort.

6.2 Exponential System Approximation for MS systems

In this section, a similar study is conducted for the MS system. Table 5 shows the input parameters ranges used in the full factorial experiment. In total 3600 experiments are conducted. As seen from the table, the experiment design considered 9 different capacity combinations ($K_1\mu_1$ and μ_2), 25 different SCV combinations ($c^2_{S,1}$ and $c^2_{S,2}$), and 16 different combinations of the population constraint (K_1 and K_2). In all the experiments, station 1 is

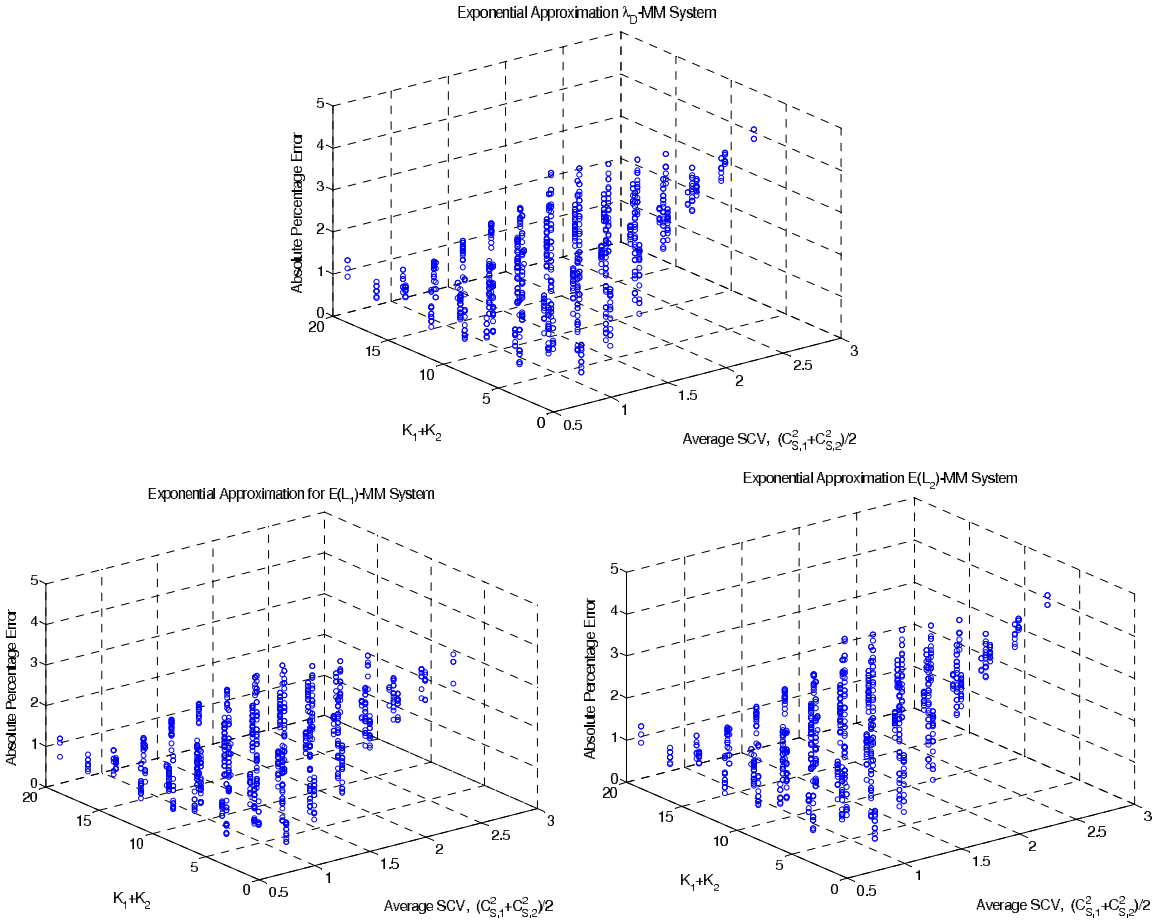


Figure 8: Performance of approximation for MM systems

the multi-server station, and station 2 is a single server station (i.e. $c_2 = 1$). Further, in all experiments, $K_1 = c_1$.

The results from the detailed experimental study are reported in Table 6. As with the MM systems, the table reports the average as well as the maximum values of the percentage errors in throughput (δ), and mean queue lengths (ϵ_1 and ϵ_2). In addition to reporting the overall errors, the table also documents how these errors vary with station capacities, number of servers, population constraints, and SCVs. From the table, the following observations are made:

(i) From the overall percentage errors reported in the table, it appears that the average percentage errors in throughput and mean queue length estimates are approximately 2%. However, unlike the MM systems, the maximum error in performance estimates could be relatively high (when compared to MM systems), being roughly 9% for the

$K_1\mu_1$	K_1	$c^2_{S,1}$	μ_2	K_2	$c^2_{S,2}$
1.25	2	0.5	1.25	2	0.5
1	5	1	1	5	1
0.8	10	1.5	0.8	10	1.5
	25	2		25	2
		2.5			2.5

Table 5: Design of experiment for MS systems

throughput and 11% for mean queue length. Interestingly, the mean queue length errors (both in terms of average and maximum) seem to be relatively less (1.5% for the average and 7.2% for the maximum) at the buffer following the multi-server station (i.e. $E(L_1)$ at buffer, B_1) when compared to the mean queue length errors at the buffer following the single server station (i.e. $E(L_2)$ at buffer, B_2) where the errors are 2.1% (average) and 11.7% (maximum). This indicates that there may be certain configurations of the MS systems where the exponential system might not provide adequately accurate performance estimates.

(ii) With respect to imbalances in station capacities, unlike the MM systems, it appears that the errors in performance estimates do not increase or decrease with station capacity imbalances. One possible explanation is that in an MS system, the station rates are unbalanced even when station capacities are balanced.

(iii) With respect to the influence of population constraint and number of servers on the percentage errors, it is evident that the error in throughput estimate decreases with increase in $K_1 + K_2$. A similar trend is observed for the mean queue length $E(L_1)$ at buffer B_1. This trend was observed for MM systems as well. However, unlike in MM systems, as $K_1 + K_2$ increases the error in estimation of mean queue length, $E(L_2)$ at buffer B_2 increases for MS systems. For instance when $K_1 + K_2$ varies from 4 to 50, the maximum error in throughput decreases from 8.1% to 2.7%; the maximum error in mean queue length, $E(L_1)$ decreases from 6.2% to 2.6%; while the maximum error in mean queue length, $E(L_2)$ increases from 5.2% to 10.2%.

(iv) Finally, with respect to the SCV of inputs, it is evident that the errors (average and maximum) in estimates of throughput and mean queue lengths increase marginally as inputs have SCVs different from 1. This trend was observed even for MM systems.

	Average			Maximum		
	δ	ϵ_1	ϵ_2	δ	ϵ_1	ϵ_2
Overall	1.757	1.467	2.101	9.054	7.195	11.665

Variation with respect station capacities

	Average			Maximum		
	δ	ϵ_1	ϵ_2	δ	ϵ_1	ϵ_2
$K_1\mu_1 = \mu_2$	2.456	2.123	0.912	9.054	7.195	3.746
$K_1\mu_1 \leq \mu_2$	1.401	1.315	2.781	8.042	7.010	11.665
$K_1\mu_1 \geq \mu_2$	1.414	0.963	2.611	7.821	5.470	10.963

Variation with respect population constraints

	Average			Maximum		
$K_1 + K_2$	δ	ϵ_1	ϵ_2	δ	ϵ_1	ϵ_2
4	3.665	2.717	1.783	8.131	6.170	5.226
7	3.034	2.434	1.898	9.054	7.195	6.708
10	2.401	2.019	2.015	6.465	5.359	5.938
12	2.125	1.794	2.181	7.598	6.497	8.900
15	1.745	1.521	2.275	5.932	5.185	8.716
20	1.303	1.167	2.438	4.705	4.187	8.430
27	1.335	1.132	1.950	6.201	4.840	11.665
30	1.174	1.011	2.096	5.180	4.849	11.276
35	0.729	0.674	2.270	3.436	3.210	10.873
50	0.459	0.436	2.041	2.712	2.558	10.230

Variation with respect SCVs

	Average			Maximum		
$0.5(c_{S,1}^2 + c_{S,2}^2)$	δ	ϵ_1	ϵ_2	δ	ϵ_1	ϵ_2
0.50	2.091	1.871	3.090	6.866	5.734	8.900
0.75	0.995	0.871	1.402	3.456	2.784	4.476
1.00	0.346	0.289	0.600	2.599	1.971	2.842
1.25	0.897	0.758	1.238	4.302	3.432	4.226
1.50	1.541	1.294	1.941	6.201	4.849	6.431
1.75	2.159	1.798	2.494	6.094	4.850	7.677
2.00	2.784	2.299	3.056	7.232	5.845	8.906
2.25	3.344	2.741	3.541	8.245	6.601	10.399
2.50	3.852	3.137	3.960	9.054	7.195	11.665

Table 6: Summary of results for MS systems

As with the MM systems, the efficiency of the exponential approximation is tested for input distribution distinct from the Coxian distribution. The performance of a fork/join system with lognormal inputs is evaluated using a simulation model built in Arena and the results

recorded for 162 scenarios. In particular, $K_1\mu_1$ is set equal to 1 and μ_2 is permitted to take values from the set (0.8, 1, 1.25). In these experiments, the parameters are chosen so that $c^2_{S,i}, i = 1, 2$ takes values from the set (0.5, 1.5, and 2.5) and $c^2_{S,1} = c^2_{S,2}$. In each case, the parameters of the lognormal distributions are chosen appropriately. In addition, K_1 takes values from the set (2, 10, 50) and K_2 takes values from the set (5, 15), and in all 162 settings, K_1 is set equal to c_1, $c_1 > 1$, and $c_2 = 1$. The results are summarized in Table 7 below. From the table it is evident that the average and maximum percentage errors are similar to those obtained for the case of Coxian inputs, suggesting that the efficiency of the exponential approximation might not depend on the particular distribution of the inputs.

Average			Maximum		
δ	ϵ_1	ϵ_2	δ	ϵ_1	ϵ_2
1.400	1.230	2.520	6.150	5.410	9.240

Table 7: Performance of approximation for MS systems with lognormal inputs

The results from these experiments indicate that for some MS system configurations, efficient approximations for performance measures can be obtained by analyzing the corresponding MS system assuming exponential inputs. The performance estimates are fairly robust across choice of distributions. However, given the relatively high values of maximum errors in performance estimates for some system configurations, it is important to identify the parameter tuple settings where such high errors could be expected. To identify specific zones where the error in approximations could be high, the percentage errors are plotted against $K_1 + K_2$ and $0.5(c^2_{S,1} + c^2_{S,2})$ in Figure 9 for throughput, λ_D, and mean queue lengths $E(L_1)$ and $E(L_2)$. It appears from these plots that in general the errors show similar trends as observed for the MM systems. However, unlike MM systems, the errors in the MS systems are higher and it is possible that in some settings, the errors from the exponential approximation are unacceptable. In these settings, exact analysis assuming Coxian inputs could be used to estimate performance. As the experiments suggest, the performance estimates demonstrate reasonable amount of insensitivity to the particular choice of the distributions, as long as the corresponding two-moments are matched.

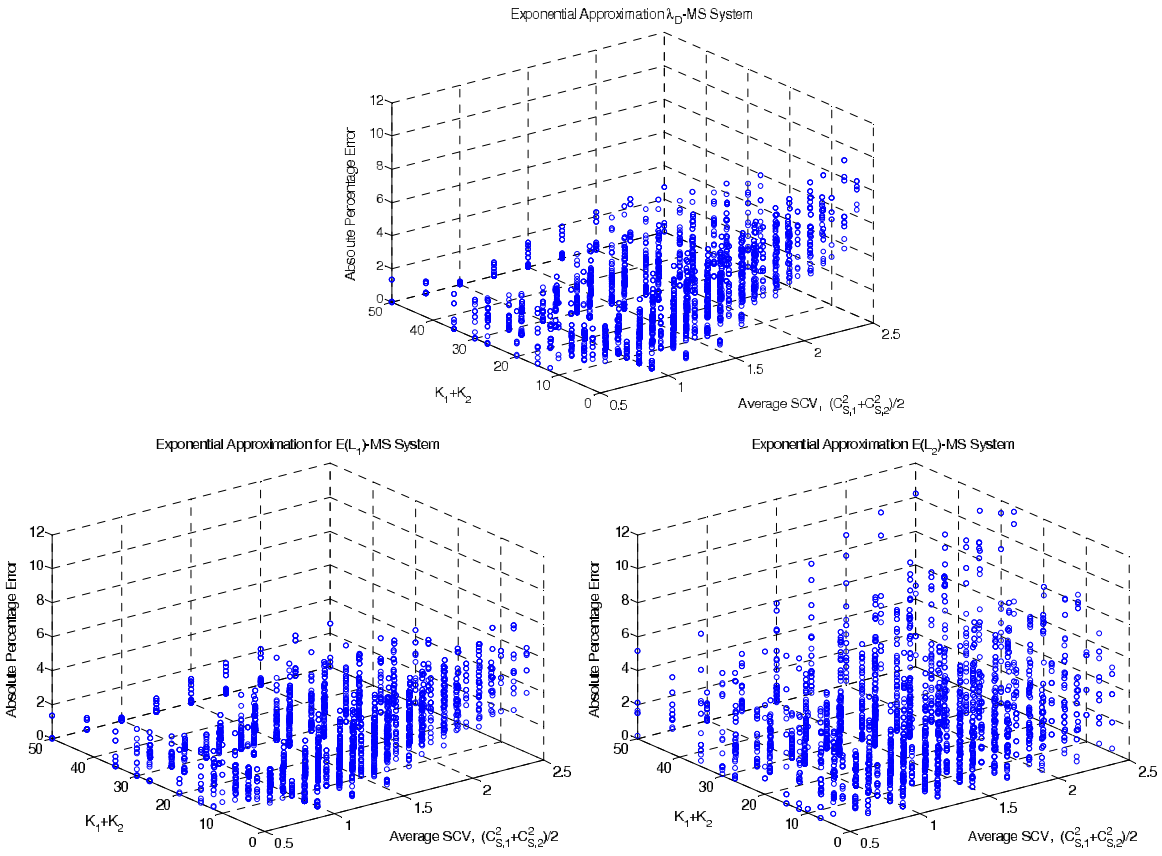


Figure 9: Performance of approximation for MS systems

7 Analysis of Variability in Inter-departure Times

This section investigates the effect of inputs from multi-server stations on the variability of inter-departure times from the fork/join station. In order to determine the exact distribution of inter-departure times, an exact analysis of the underlying semi-Markov process will need to be conducted. Instead of conducting this detailed analysis, simulation experiments are conducted to empirically determine the distribution of the inter-departure times from the fork/join stations. In particular, the SCV of inter-departure time distribution, c_D^2, is recorded for several MM, MS and SS systems with different SCVs of inputs.

Sample results from the simulation study conducted using Arena are reported in the Table 8 below. For all systems (SS, MM and MS), the experiment considered configurations with balanced and unbalanced station capacities. For the SS system, station capacities are defined by $\mu_i, i = 1, 2$, while for the MM system, station capacities are defined by $K_i\mu_i, i = 1, 2$. For the MS system, station 1 is a multi-server station with $c_1 = K_1$, and its station capacity is

defined by $K_1\mu_1$, while for the single server station, station 2, $c_2 = 1$ its capacity is defined by μ_2. Further, in an MS system by suitably defining station capacities, one could consider systems where either station 1 or station 2 has the smaller capacity. In the experiment, both cases were considered. The population constraint was varied from 10 to 50, and input SCVs, $c^2_{S,i}, i = 1, 2$ with values smaller and larger than 1 where considered.

For the SS systems, the following observations are made: (i) For a system with balanced capacities (i.e. $\mu_1 = \mu_2$), the SCV of the inter-departure times tends to the average of the SCVs of the service times of the servers at the two stations with increase in $K_1 + K_2$. (ii) For a system with unbalanced capacities (i.e. $\mu_1 \neq \mu_2$), with increase in $K_1 + K_2$, the SCV of the inter-departure times tends to the SCV of the service times of the servers having the smaller mean service time.

For the MM systems, the following observations are made: (i) Unlike the SS systems, for the MM systems, with increase in $K_1 + K_2$, the SCV of the inter-departure times c^2_D tends to 1 regardless of whether station capacities are balanced or unbalanced. Note that since in the MM systems, $K_i = c_i, i = 1, 2$, increase in $K_1 + K_2$ implies that the $c_1 + c_2$ increases as well. This suggests that the inter-departure time distribution from the MM system might be exponentially distributed. (ii) The graphs in Figure 10 compares the plot of the cumulative distribution function of the inter-departure times drawn using data obtained from simulation with the cumulative distribution function of an exponential random variable with the same mean. In the figures, $R = \frac{c_1\mu_1}{c_2\mu_2}$. In Figure 10 (a) $K_1\mu_1 = 1$ $K_2\mu_2 = 1.25$, $c^2_{S,1} = 2.5$, $c^2_{S,2} = 0.5$, $K_1 = K_2 = 10$. In Figure 10 (b) the parameters are the same as in Figure 10 (a), except that $K_1 = K_2 = 50$. In Figure 10 (c) $K_1\mu_1 = K_2\mu_2 = 1$, $c^2_{S,1} = 2.5$, $c^2_{S,2} = 0.5$, $K_1 = K_2 = 10$. Similarly, in Figure 10 (d) the parameters are the same as in Figure 10 (c), except that $K_1 = K_2 = 50$. As can be observed from the plot, the true inter-departure time distribution appears to be very close to an exponential distribution. Recall that for open $GI/G/c$ queues, the departure process tends to a Poisson process as the mean number of busy servers increases [32]. It is conjectured that in fork/join stations with inputs from multi-server stations, a similar phenomenon might hold. A formal proof of this conjecture will require a detailed analysis of the departure process and is part of future research.

SS systems

Capacity Station 1	Capacity Station 2	(K_1, K_2)	$c^2_{S,1}$	$c^2_{S,2}$	λ_D^{-1}	c^2_D
1	1	(10,10)	2.5	0.5	1.070	1.431
1	1	(50,50)	2.5	0.5	1.010	1.532
1	1	(10,10)	2.5	2.5	1.110	2.346
1	1	(50,50)	2.5	2.5	1.020	2.491
1	1.25	(10,10)	2.5	0.5	1.010	2.176
1	1.25	(50,50)	2.5	0.5	1.000	2.496
1	1.25	(10,10)	2.5	2.5	1.030	2.383
1	1.25	(50,50)	2.5	2.5	1.000	2.496

MM systems

Capacity Station 1	Capacity Station 2	(K_1, K_2)	$c^2_{S,1}$	$c^2_{S,2}$	λ_D^{-1}	c^2_D
1	1	(10,10)	2.5	0.5	1.140	0.865
1	1	(50,50)	2.5	0.5	1.060	0.944
1	1	(10,10)	2.5	2.5	1.156	0.989
1	1	(50,50)	2.5	2.5	1.065	0.973
1	1.25	(10,10)	2.5	0.5	1.055	0.96
1	1.25	(50,50)	2.5	0.5	1.006	0.999
1	1.25	(10,10)	2.5	2.5	1.062	1.017
1	1.25	(50,50)	2.5	2.5	1.006	1.004

MS systems

Capacity Station 1	Capacity Station 2	(K_1, K_2)	$c^2_{S,1}$	$c^2_{S,2}$	λ_D^{-1}	c^2_D
1	1	(10,10)	2.5	0.5	1.081	0.876
1	1	(50,10)	2.5	0.5	1.058	0.744
1	1	(10,10)	2.5	2.5	1.124	1.524
1	1	(50,10)	2.5	2.5	1.080	1.722
1	0.8	(10,10)	2.5	0.5	1.265	0.629
1	0.8	(50,10)	2.5	0.5	1.252	0.517
1	1.25	(10,10)	2.5	0.5	1.014	1.07
1	1.25	(50,10)	2.5	0.5	1.005	0.982

Table 8: Comparison of mean and SCV of inter-departure times

For the MS systems, the following observations are made: (i) For systems where station capacities are balanced, no clear trends are observed with respect to the behavior of the SCV of inter-departure times, c^2_D with increase in $K_1 + K_2$. Depending on the system configuration and input parameters, the SCV value might either tend towards the average of the SCV of the two inputs (as in SS systems), or towards 1 (as in MM systems). (ii) For systems where

station capacities are unbalanced, and the single server station has smaller station capacity, with increase in $K_1 + K_2$, the SCV of inter-departure times tends towards the SCV of the service times of this station. (iii) For systems where station capacities are unbalanced, and the multi-server station has smaller station capacity, with increase in $K_1 + K_2$, the SCV of inter-departure times tends towards 1.

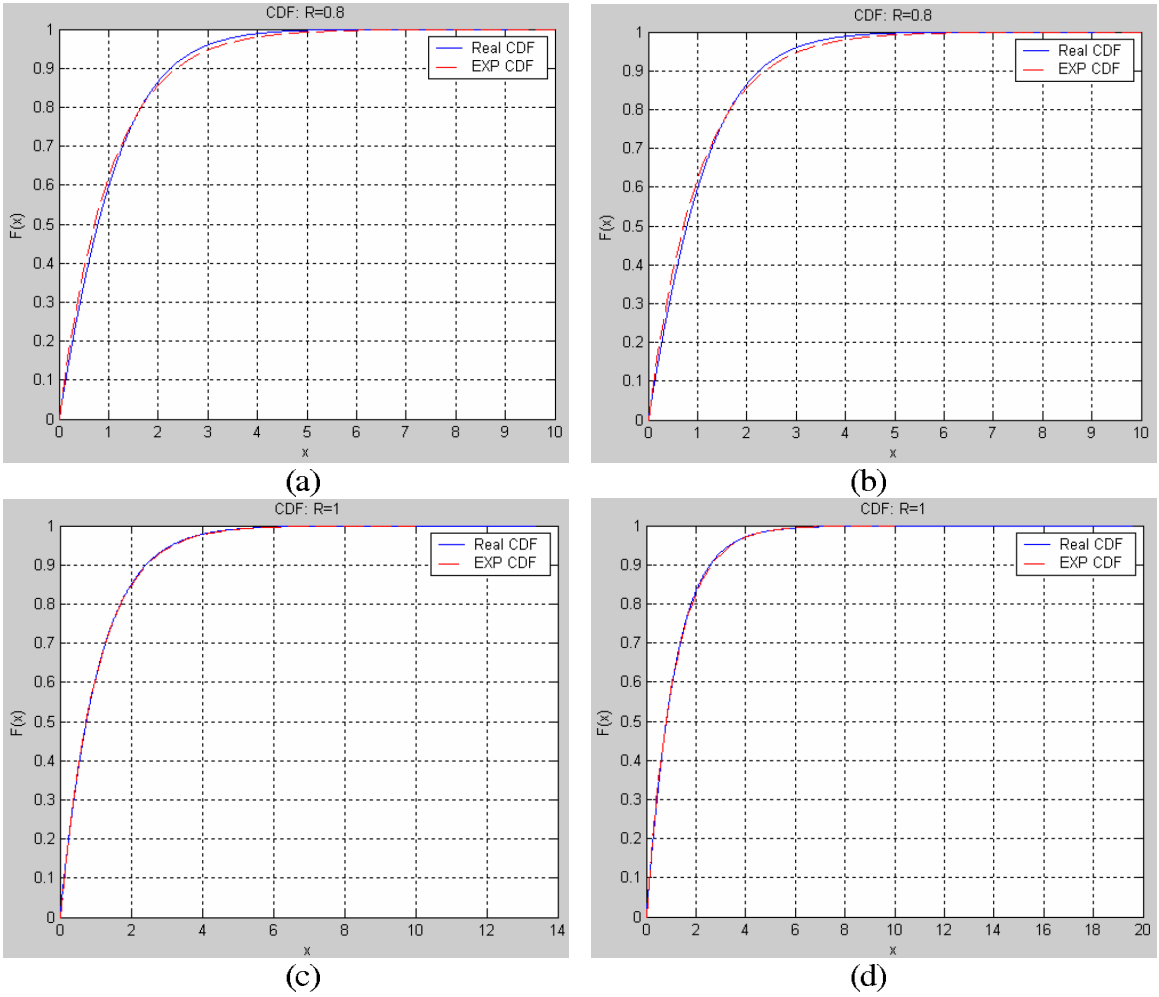


Figure 10: Distribution of inter-departure times from MM systems

8 Conclusions

This paper presents an exact analysis of a fork/join station in a closed queuing network with inputs from stations composed of multiple servers with two-phase Coxian distributions. The choice of the two-phase Coxian distribution permits the analysis of more general input pro-

cesses without much added computational complexity. Using an exact analysis of the queue length process, it is shown that when the number of servers at the two stations is large, for some system configurations, the variabilities in the input processes have a negligible effect on throughput and queue length distributions for MM systems and MS systems. Such insights are in contrast to known results for fork/join stations with inputs from single server stations (SS systems), where variability has been known to significantly influence system performance. These differences can be highlighted using the explicit models developed in this research for fork/join stations with inputs from multi-server stations.

The relative insensitivity of system performance to variability inputs in certain settings, suggests the possibility that exact results from the analysis of the simpler system with exponential inputs could provide efficient approximations for performance measures. Extensive experiments are conducted to precisely quantify regions in the parameter design space where such approximations will be efficient. Finally some insights with respect to the variability of departure process from the fork/join station are also provided. These insights could be used directly in the design of systems containing fork/join stations. In addition, when the fork/join station is part of a larger network, the suggested approximations could permit efficient analysis of larger systems without compromising on the accuracy of performance estimation.

Finally, the analysis here is also useful in extending decomposition methods for performance analysis of queuing networks. Such methods require efficient "two-moment" approximations to characterize the performance of stations in the network. However, such approximations were previously not available for fork/join stations with inputs from multi-server stations. The analysis and insights in this paper have been used in the development and validation of two-moment approximations for fork/join stations (Goossens[10]). These approximations have been used to develop decomposition methods for the analysis of blocking phenomenon in multi-server tandem lines with no buffers.

Acknowledgements: The research of the first author was supported by a grant provided by the Research Foundation - Flanders (FWO) and was conducted while he was a graduate student at University of Antwerp, in the Faculty of Applied Economics, Environment, Technology and Technology Management Department.

Bibliography

[1] T. Altiok, *Performance Analysis of Manufacturing Systems*, Springer Series in Operations Research, New York, New York, (1996).

[2] F. Baccelli and A.M. Makowski, *Queueing models for systems with synchronization constraints*, Proceedings of the IEEE **77** (1989), no. 1, 138–161.

[3] F. Baccelli, W.A. Massey, and D. Towsley, *Acyclic fork/join queuing networks*, Journal of the ACM **36** (1989), 615–642.

[4] B. Baynat and Y. Dallery, *Approximate techniques for general closed queueing networks with subnetworks having population constraints*, European Journal of Operational Research **69** (1993), 250–264.

[5] U.N. Bhat, *Finite capacity assembly like queues*, Queueing Systems **1** (1986), 185–201.

[6] F. Bonomi, *An approximate analysis for a class of assembly-like queues*, Queueing Systems **1** (1987), 289–309.

[7] J. A. Buzacott and J. G. Shantikumar, *Stochastic Models of Manufacturing Systems*, Prentice Hall, Englewood Cliffs, New Jersey, (1993).

[8] Liesje de Boeck, *Performance analysis of assemblies in open queueing systems*, Ph. D Thesis, Faculty of Applied Economics, University of Antwerp (2003).

[9] M. Di Mascolo, Y. Frein, and Y. Dallery, *An analytical method for performance evaluation of kanban controlled production systems*, Operations Research **44** (1996), no. 1, 50–64.

[10] N. Goossens, *Blocking systems with zero buffer: A kanban-like approach*, Ph. D Thesis, Faculty of Applied Economics, University of Antwerp (2006).

[11] N. Goossens, A. Krishnamurthy, and N. Vandaele, *Analysis of a fork/join station with inputs from a finite population sub-network with multi-server stations*, Queuing Systems – under review (2006).

[12] J.H. Harrison, *Assembly-like queues*, Journal of Applied Probability **10** (1973), no. 2, 354–367.

[13] W.J. Hopp and J.T. Simon, *Bounds and heuristics for assembly-like queues*, Queueing systems **4** (1989), no. 2, 137–156.

[14] C. Knessl, *On the diffusion approximation to a fork and join queueing model*, SIAM Journal of Applied Mathematics **51** (1997), no. 1, 160–171.

[15] S.-S. Ko and R.F. Serfozo, *Response times in m/m/s fork-join networks*, Advances in Applied Probability **36** (2004), 854–871.

[16] A. Krishnamurthy and R. Suri, *Performance analysis of single stage kanban controlled production systems using parametric decomposition*, Queueing Systems **54** (2006), 141–162.

[17] A. Krishnamurthy, R. Suri, and M. Vernon, *Two-moment approximations for throughput and mean queue length of a fork/join station with general input processes*, Stochastic modeling and optimization of manufacturing systems and supply chains, Shanthikumar, J.G., Yao D.D., and Zijm, W.H.M. (Editors), International Series in Operations Research and Management Science, Kluwer Academic Publishers (2003), 87–126.

[18] A. Krishnamurthy, R. Suri, and M. Vernon, *Analysis of a fork/join synchronization station with inputs from coxian servers in a closed queuing network*, Annals of Operations Research **125** (2004), 69–94.

[19] A. Kumar and R. Shorey, *Performance analysis and scheduling of stochastic fork-join jobs in a multicomputer system*, IEEE Transactions on Parallel and Distributed Systems **4** (1993), no. 10, 1147–1164.

[20] G. Latouche, *Queues with paired customers*, Journal of Applied Probability **18** (1981), no. 3, 684–696.

[21] Y.C. Liu and H.G. Perros, *Approximate analysis of a closed fork/join model*, European Journal of Operational Research **53** (1991), no. 3, 382–392.

[22] R. Nelson and A.N. Tantawi, *Approximate analysis of fork/join synchronization in parallel queues*, IEEE Transactions on Computers **37** (1988), no. 6, 739–743.

[23] B. Prabhakar, N. Bambos, and T.S. Mountford, *The synchornization of poisson processes and queueing networks with service and synchronization nodes*, Advances in Applied Probability **32** (2000), 824–843.

[24] P.C. Rao and R. Suri, *Approximate queueing models of fabrication / assembly systems: Part i-single level systems*, Production an Operations Management **3** (1994), 244–275.

[25] R. Ramakrishnan and A. Krishnamurthy, *Analytical Approximations for Kitting Systems with Multiple Inputs*, Asia-Pacific Journal of Operations Research (2007), To Appear.

[26] P.C. Rao and R. Suri, *Performance analysis of an assembly station with input from multiple fabrication lines*, Production and Operations Management **9** (2000), no. 3, 283–302.

[27] P. Som, W.E. Wilhelm, and R.L. Disney, *Kitting process in a stochastic assembly system*, Queueing Systems **17** (1994), 471–490.

[28] M. Takahashi, H. Osawa, and T. Fujisawa, *A stochastic assembly system with resume levels*, Asia-Pacific Journal of Operational Research **15** (1998), 127–146.

[29] M. Takahashi, H. Osawa, and T. Fujisawa, *On a synchornization queue with two finite buffers*, Queueing Systems **36** (2000), 107–123.

[30] E. Varki, *Mean value technique for closed fork-join networks*, Proceedings of ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems (1999).

[31] S. Varma and A. Makowski, *Interpolation approximations for symmetric fork-join queues*, Performance Evaluation **20** (1994), 245–265.

[32] W. Whitt, *Approximations for the GI/G/m queue*, Production and Operations Management **2** (1993), 114–161.